AD-A057 375 CALIFORNIA UNIV LOS ANGELES DEPT OF SYSTEM SCIENCE F/G 17/2
CODING AND MODULATION FOR COMMUNICATION CHANNELS WITH MEMORY.(U)
MAR 78 E BIGLIERI N00014-76-C-0875
UNCLASSIFIED UCLA-ENG-7829 NL

1 OF 2
AD
A057375

LEVEL II

This work was supported by the
U.S. Office of Naval Research
under Contract ~~ONR-P-0030-N-76~~

not a contract per phone conversation 18 Aug 78

UCLA-ENG-7829
MARCH 1978

# CODING AND MODULATION FOR COMMUNICATION CHANNELS WITH MEMORY

EZIO BIGLIERI

UCLA • SCHOOL OF ENGINEERING AND APPLIED SCIENCE

UCLA:ENG-7829
March 1978

# CODING AND MODULATION

# FOR COMMUNICATION CHANNELS WITH MEMORY

Ezio Biglieri
Istituto Elettrotecnico
Università di Napoli (Italy)

and

Department of System Science
University of California, Los Angeles

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. ~~REPORT NUMBER~~ 2. ~~GOVT ACCESSION NO.~~ 3. ~~RECIPIENT'S CATALOG NUMBER~~

9 Technical rept. Mar 77-Mar 78,

4. ~~TITLE (and Subtitle)~~

6 CODING AND MODULATION FOR COMMUNICATION CHANNELS WITH MEMORY.

5. ~~TYPE OF REPORT~~ & PERIOD COVERED

Tech. Report: Mar 77-Mar 78

14 ~~6. PERFORMING ORG. REPORT~~ NUMBER

UCLA-ENG-7829

7. ~~AUTHOR(s)~~

10 Ezio Biglieri
Istituto Elettrotecnico
Università di Napoli (Italy)

15 ~~8. CONTRACT OR GRANT N~~UMBER(s)

N00014-76-C-0875

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Department of System Science
University of California
Los Angeles, CA 90024

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

12 138p.

11. CONTROLLING OFFICE NAME AND ADDRESS

Office of Naval Research
Electronics Program, Code 427
Arlington, VA 22217

11 ~~12. REPORT DATE~~

Mar [redacted] 78

~~13. NUMBER OF PAGES~~

134

14. MONITORING AGENCY NAME & ADDRESS (*if different from Controlling Office*)

15. SECURITY CLASS. (*of this report*)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

Coding and modulation -- channels with memory -- sequential machines -- offset PSK signals -- computation of power spectrum -- error probability of digital communication -- nonlinear channels

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

Linear and nonlinear digital communication channels with memory are considered, with the goal of providing a number of general tools for the analysis and design of coding and modulation schemes operating on such channels.

405 213

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

## TABLE OF CONTENTS

## ABSTRACT

Linear and nonlinear digital communication channels with memory are considered, with the goal of providing a number of general tools for the analysis and design of coding and modulation schemes operating on such channels.

## *FOREWORD - MOTIVATION AND SUMMARY OF THE WORK*

The increasing demand for high-speed digital communication has resulted in new theoretical problems for communication engineers. In particular, analytical tools are called for that allow the performance of digital transmission systems to be evaluated accurately, and coding or modulation schemes to be designed properly.

The channel model that will be dealt with in this Report is a bandlimited one -- possibly nonlinear. In fact, in most digital communication systems, for an efficient use of the frequency, only a restricted bandwidth is available. Moreover, in some cases, such as satellite repeaters, amplifiers are working at or near saturation for better efficiency, so that the combined effects of bandlimiting and nonlinearity must be taken into account for a proper analysis of the system.

The purpose of this Report is to gather a number of analytical tools that have been devised for designing or analyzing coding and modulation schemes for digital bandlimited communication channels, i.e., channels with memory.

Chapter 1 is devoted to the description of some coding and modulation schemes devised for real-life channels. Essentially, two philosophies are involved in those schemes. The first is intended to restrict the symbol stream to form only sequences that are well matched to the channel features. The second is to constrain the power spectrum of transmitted signals in order to concentrate it in spectral regions where the channel response is better.

Design techniques based on both philosophies are considered in Chapter 2.

In Chapter 3, the problem of characterizing continuous nonlinear channels with memory is considered. A Volterra-series approach is taken and a method is proposed for computing the error probability at the output of such channels.

Finally, in Chapter 4, a Markov-chain technique is proposed for modeling discrete-time, nonlinear channels with memory. This model can be used for several applications: among them, I shall consider the computation of power spectra at the channel output, the derivation of maximum-likelihood sequence demodulators and the evaluation of error statistics.

# CHAPTER 1 - SOME EXAMPLES OF CODING AND MODULATION SCHEMES FOR CHANNELS WITH MEMORY

## 1.1 INTRODUCTION

In this chapter, I shall describe some examples of coder-modulator pairs devised, and sometimes actually used, for transmitting digital information over channels with memory. This is intended to provide motivation for the material of later Chapters; there, some general tools will be developed that will allow these transmission schemes to be analyzed for the purpose of deriving general performance indices, such as spectral occupancy or error probability.

The transmission scheme with which we shall be dealing is not uncommon in Information Theory (see Fig. 1.1). We will assume that the information source generates a sequence of B-ary symbols $a_n$, $-\infty < n < \infty$ , which are equally likely and statistically independent of each other.

The information sequence is sent to a coder, which maps it into a sequence of M-ary symbols $b_n$ , and then to a modulator. We shall assume that the modulator is memoryless, i.e., it performs a one-to-many mapping of coded M-ary symbols into a set of waveforms of equal duration T (say).

For generality, we shall allow coded symbols and modulator waveforms to be either real or complex. In particular, symbols and waveforms will be assumed to be real when the transmission channel is baseband, and complex when the transmission is bandpass. This convention is based on the use of complex envelopes to represent bandpass signals and systems.

The pioneering work of Nyquist (1924, 1928) was devoted to optimizing the modulation process, i.e., the choice of waveforms that allow the effects of the channel memory to be coped with when the channel is assumed to be bandlimited, linear and perfectly known.

More recently, it has been recognized that coding can be used to improve the performance of a data communication system. Codes can be used to provide protection against certain types of errors that may occur during transmission; however, most of the results of coding theory are based on a memoryless channel assumption. Although it has been experimentally recognized that some codes designed to operate on a memoryless channel can improve the performance of a channel with memory, no general results are available in the area of coding for channels with memory.

Coding schemes have been developed for use in pulse-code modulation (PCM) systems, high-speed data communication systems, and high-density magnetic recording. As we shall see more thoroughly in Chapters 1 and 2, these schemes are designed to cope with unwanted features of the channel. For example, if the channel exhibits a poor frequency response in some regions, one may want the transmitted signal to contain relatively little power in the imperfect regions of the channel. In this way, we can reasonably expect that the amount of distortion produced by the channel will be reduced with respect to the case of no coding.

A slightly different philosophy leads to the concept of codes designed in the time domain. The basic idea is to avoid transmitting certain symbol sequences that are not well matched to the channel characteristics. For example, sequences giving rise to a constant, non-zero voltage on a transmission channel must be avoided if the channel has a poor response around the zero frequency.

In the next section, we shall consider some common examples of binary (M=2), ternary (M=3) and multilevel (M>3) codes designed for

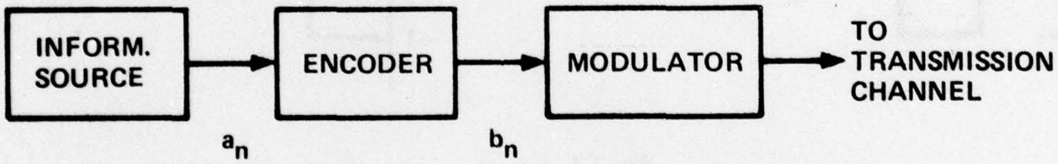

INFORM. SOURCE
ENCODER
MODULATOR
TO TRANSMISSION CHANNEL
$a_n$
$b_n$

Fig. 1.1

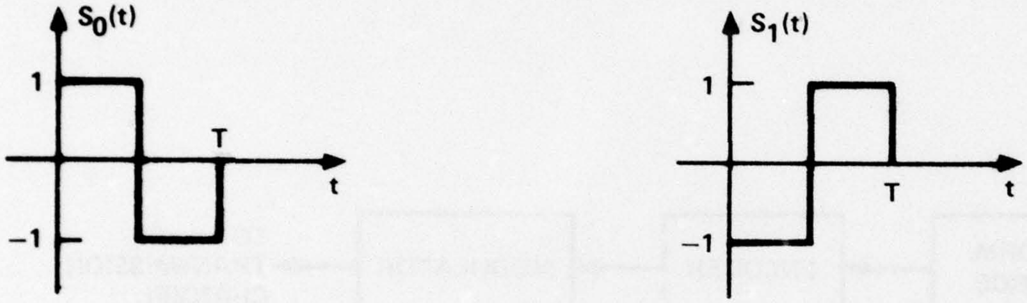


Fig. 1.2

transmission of digital data. For other examples see Croisier (1970) and Kobayashi (1970 and 1971), as well as the references therein.

## 1.2 BINARY CODES

The simplest way to transmit binary digital data on a line is to associate two different waveforms, say $s_0(t)$ and $s_1(t)$ , to source output symbols "0" and "1" . Since generally the medium to be used for transmission has a poor dc response, one wants to avoid transmitting a dc term. If source zeros and ones are equally likely, the transmitted signal is dc-free if

$$s_1(t) = - s_0(t) \tag{1.1}$$

A common signaling format for baseband signals uses positive and negative rectangular pulses with duration $T$ , where $T^{-1}$ is the rate, in binary symbols/sec , at which the source outputs its symbols. In the simplest format, called non-return-to-zero (NRZ), a source "one" is thus represented by one voltage level, and a "zero" by its negative. Transmission of dc is thus avoided, but the power spectrum of this signal format is concentrated around the zero frequency, a spectral zone where the response of the transmission medium is usually poor.

A better distribution of signal energy can be obtained using two different waveforms $s_0(t)$ and $s_1(t)$ , i.e., using a different modulation scheme. A possible choice is depicted in Fig. 1.2 . With this choice (Split-phase, or Manchester code, or binary phase modulation), the power of the transmitted signal is concentrated toward higher frequencies, where the medium response is expected to be better.

A more sophisticated solution can be achieved by coding the source output. An interesting example of such a scheme, known as Delay Modulation or Miller Code (Hecht and Guida, 1969), can be modeled as follows: consider the 4-state sequential machine, driven by source outputs, depicted in Fig. 1.3 .

To each of the four states, a waveform is associated according to the rule summarized in Fig. 1.4 . Every source symbol output produces a transition from one state to another (Fig. 1.3); an example of the waveform at the output of the modulator is given in Fig. 1.5 .

As a result, with this code, the signal power spectrum is highly concentrated around frequencies less than .5/T ; moreover, the spectrum is small in the vicinity of the zero frequency. This code is actually used for magnetic tape recording; it is also attractive for phase-shift-keyed signaling (see Lindsey and Simon, 1973, p. 11).

Binary block codes with frequency spectrum constraints

Gorog (1968) has analyzed families of binary block codes that are both redundant (i.e., error-detecting) and have the signal energy concentrated into a predetermined range of the frequency spectrum.

Consider a sequence of $N$ coded symbols $b_0, b_1, \ldots, b_{N-1}$ , and assume that each symbol, taking the value $\pm 1$ , amplitude-modulates a basic waveform $s(t)$ with duration $T$ . The signal at the output of the modulator is

$$x(t) = \sum_{i=0}^{N-1} b_i \, s(t-iT) \qquad (1.2)$$

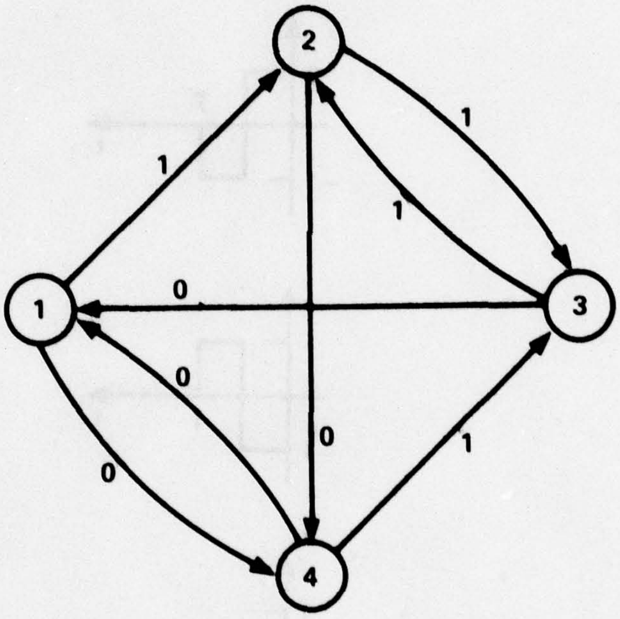


Fig. 1.3

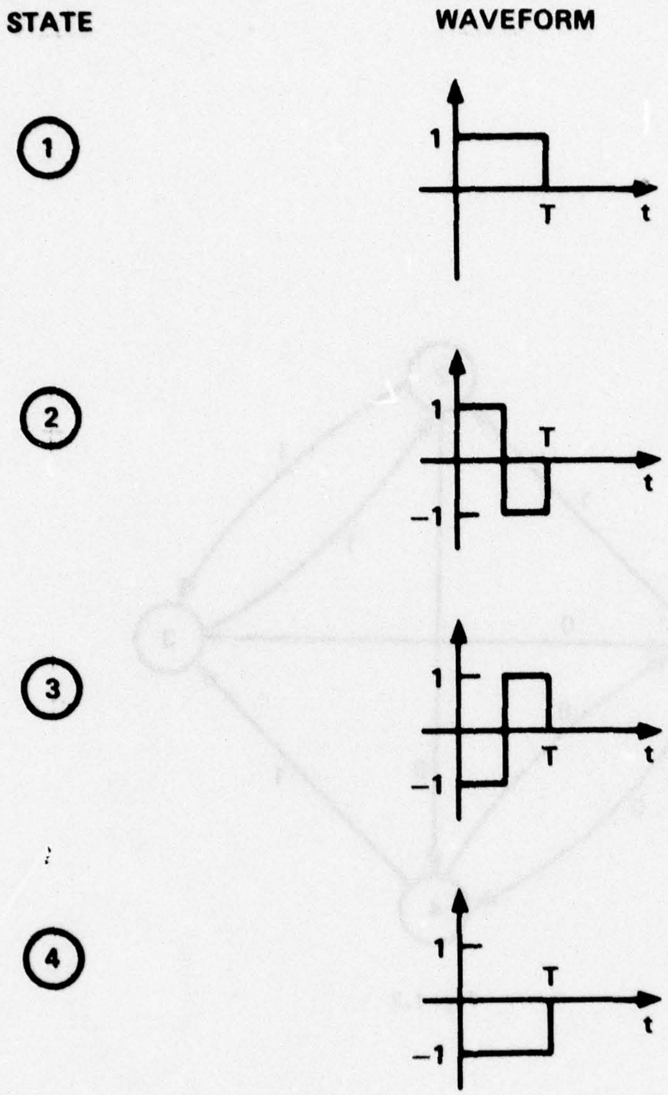


Fig. 1.4

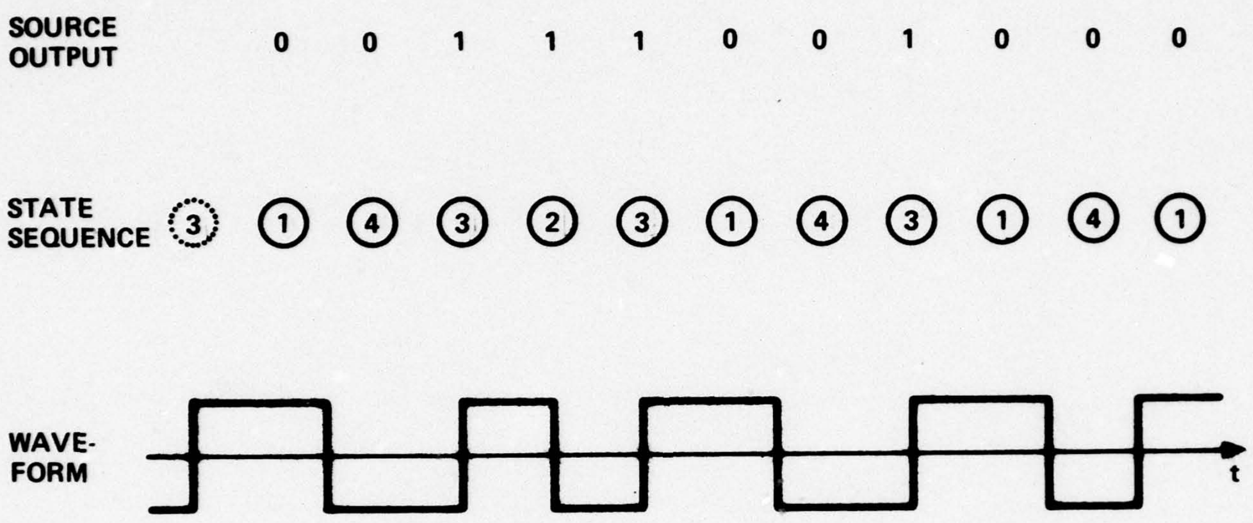

SOURCE OUTPUT
0 0 1 1 1 0 0 1 0 0 0
STATE SEQUENCE
3 1 4 3 2 3 1 4 3 1 4 1
WAVE-FORM
t

Fig. 1.5.

and its Fourier transform is

$$X(\omega) = S(\omega) \sum_{i=0}^{N-1} b_i e^{-ij\omega T} , \qquad j=\sqrt{-1} \tag{1.3}$$

where $S(\omega)$ is the Fourier transform of $s(t)$ .

As can be seen, the code contributes to the spectrum through the factor

$$C(\omega) \overset{\Delta}{=} \sum_{i=0}^{N-1} b_i e^{-ij\omega T} \tag{1.4}$$

We shall now see how to construct codes that give rise to zeros occurring at $\omega=0$ or $\omega=\pi/T$ , which are frequency values where channel imperfections usually occur.

Consider first $\omega=0$ , i.e., the coded sequences whose spectrum is zero at the origin. A code of this kind is useful for transmission over channels with poor response at low frequencies.

To get $C(0)=0$ , the condition

$$\sum_{i=0}^{N-1} b_i = 0 \tag{1.5}$$

must hold; i.e., the coded sequence must contain the same number of $+1$'s as $-1$'s . If $N=k\nu$ , the sequence can be decomposed into $k$ subsequences of $\nu$ symbols each. Thus, a sufficient condition for (1.5) to be satisfied is that

$$\sum_{i=(j-1)\nu}^{j\nu-1} b_i = 0 \qquad j=1,2,\ldots,k \tag{1.6}$$

that is, the j-th subsequence must contain the same number of +1's and -1's (this requires, of course, that $\nu$ be even).

If each subsequence is a codeword, the code has no more than

$$\binom{\nu}{\nu/2}$$

codewords.

Another constraint that can be imposed is that $X(\omega)$ be zero at $\omega=\pi/T$. This can be obtained if $C(\pi/T) = 0$ ; i.e.,

$$\sum_{i=0}^{N-1} b_i(-1)^i = 0 \tag{1.7}$$

A sufficient condition for (1.7) to be satisfied is

$$\sum_{i=(j-1)\nu}^{j\nu-1} a_i(-1)^i = 0 \qquad j=1,2,\ldots,k \tag{1.8}$$

For $\nu$ even, a subsequence will satisfy this condition if a reversal of the signs of its odd-numbered bits produces a subsequence with the same number of +1's as -1's . There are

$$\binom{\nu}{\nu/2}$$

different subsequences of length $\nu$ that meet this condition.

Gorog (1968) also gives solutions to the case when $\nu$ is odd, and when the constraint that $C(0) = C(\pi/T) = 0$ is imposed. For example, in the case $\nu=8$ . there are 36 codewords such that the spectrum of the sequence will have a zero at $\omega=0$ and $\omega=\pi/T$ . The resulting code is represented in Table 1.1; notice that the number of codewords (36) makes this code useful for transmission of alphanumeric symbols (10 numbers + 26 letters).

## 1.3 TERNARY CODES

A certain amount of transmission power can be saved if, within the set of waveforms used by the modulator, one wants to include a zero waveform. In a binary code, the simplest such format is the so-called unipolar format: the binary symbols 1 and 0 , left uncoded, are transmitted as presence and absence of pulses, respectively.

*With this format, long sequences of 1's can occur which result in* dc wander, since the repeaters along the line are not dc-coupled to the cable medium. Thus, one can use three waveforms in the modulator: the zero waveform and two non-zero waveforms with equal shape and opposite polarities.

One of the simplest examples is the bipolar code (Aaron,1952) used in Bell System's T1 carrier PCM system. In bipolar, the source binary symbol 0 is represented by no signal on the line, and the binary symbol 1 is represented alternately by positive and negative pulses. The effects of dc wander are reduced, since a pulse of one polarity is certainly followed by a pulse of the opposite polarity.

Codes with $B=2$ and $M=3$ are often referred to as pseudoternary (PT) codes. The bipolar format is a simple example of pseudoternary code; more sophisticated schemes of such codes will be described later on (see also, Croisier(1970), and references therein).

## Linear ternary codes

*A ternary code is called* *linear time-invariant* *if the coded symbols* can be derived from the binary source symbols through the linear relation

$$b_n = \Gamma(D)a_n \qquad (D \equiv \text{delay operator}) \qquad (1.9)$$

where $\Gamma(D)$ is a polynomial that defines the code, while $a_n = \pm 1$ according to the source output.

In order for the $b_n$ to have only three possible values for any sequence of source symbols, it is necessary that $\Gamma(D)$ have only two non-zero coefficients, and that they be either equal or opposite. We have three basic linear PT codes:

| | |
|---|---|
| Duobinary: | $\Gamma(D) = 1+D$ |
| Twinned binary: | $\Gamma(D) = 1-D$ |
| Class-IV partial response: | $\Gamma(D) = 1-D^2$ |

Duobinary code is characterized by the absence of direct transitions between levels + and -, but is subject to unwanted dc wander. Twinned binary and class-IV partial-response code are free of dc.

Linear codes usually suffer a common drawback: in either case, the erroneous reception of one ternary element causes a possibly infinite error propagation. In fact, $a_n$ can be recovered from $b_n$ according to the rule:

$$a_n = \frac{1}{\Gamma(D)}\, b_n \qquad (1.10)$$

TABLE 1.1

Gorog's alphanumeric code of length 8

| Alpha-numeric | Binary Code | Alpha-numeric | Binary Code |
|---|---|---|---|
| 0 | + + - - + + - - | I | + + - - - + + - |
| 1 | - + + - - + + - | J | - + - + + - + - |
| 2 | - - + + - - + + | K | - - + - + + - + |
| 3 | + - - + + - - + | L | + - - + - + + - |
| 4 | + + + - - + - - | M | - + - - + - + + |
| 5 | - + + + - - + - | N | - + + - - - + + |
| 6 | - - + + + - - + | O | + - + + - - - + |
| 7 | + - - + + + - - | P | + + + - - - - + |
| 8 | - + - - + + + - | Q | + + - - - - + + |
| 9 | - - + - - + + + | R | + + - + + - - - |
| A | + - - + - - + + | S | - + + - + + - - |
| B | + - + - - + - + | T | - - + + - + + - |
| C | + + - + - - + - | U | + - - - - + + + |
| D | + + - - + - - + | V | - - - - + + + + |
| E | - - - + + - + + | W | - - - + + + + - |
| F | - + + - + - - + | X | - - + + + + - - |
| G | + - + + - + - - | Y | - + + + + - - - |
| H | + - - - + + - + | Z | + + + + - - - - |

so that, if for example $\Gamma(D) = 1+D$ , we get

$$a_n = \frac{1}{1-D}\ b_n$$

$$= (1+D+D^2 + \ldots)b_n$$

$$= b_n + b_{n-1} + b_{n-2} + \cdots \qquad (1.11)$$

So, if one of the $b_n$'s is received erroneously, the error affects all the subsequent $a_n$'s . To avoid this catastrophic situation, the sequence at the output of the source is first precoded according to the rule

$$a'_n \equiv \frac{1}{\Gamma(D)}\ a_n \qquad \text{mod } 2 \qquad (1.12)$$

and then coded:

$$b_n = \Gamma(D)\ a'_n \qquad (1.13)$$

In this scheme, $a_n$ can be recovered from $b_n$ through the following operation:

$$a_n \equiv \Gamma(D)a'_n \qquad \text{mod } 2 \qquad (1.14)$$

that is,

$$a_n \equiv b_n \qquad \text{mod } 2 \qquad (1.15)$$

and error propagation is avoided. It can be shown that, if $(a_n)_{n=-\infty}^{\infty}$ is a sequence of IID binary random variables, $(a'_n)_{n=-\infty}^{\infty}$ is again a sequence of IID binary random variables, so that the precoding operation does not alter source statistics.

For example, if $\Gamma(D)= 1-D$ , the precoding operation is

$$a'_n = a_n \oplus a'_{n-1} \tag{1.16}$$

where $\oplus$ denotes module-2 addition.

It can be seen that precoded twinning binary is equivalent to the bipolar code.

## Nonlinear ternary codes

We shall now describe some of the ternary codes that have a practical significance in data communications. Following Croisier (1970), we shall distinguish between alphabetic and nonalphabetic codes.

### A. Alphabetic ternary codes

Source binary data are first framed into blocks, and each is encoded in a ternary block. The block length can be kept constant, or is a variable. For constant-length codes, with binary-block length $K$ and ternary-block length $N$ , there are $2^K$ binary blocks and $3^N$ ternary blocks; so we must have

$$2^K \geq 3^N \tag{1.17}$$

or

$$K \leq log_2 3\ N \qquad (log_2 3 \simeq 1.58) \tag{1.18}$$

We see that alphabetic ternary codes may potentially increase the data rate up to 58% over uncoded binary.

Pair-selected ternary (Sipress,1965)

A widely known code of this type is Paired Selected Ternary (PST). In this code, the incoming binary sequence is framed into blocks of length K=2 , and each binary block is coded according to the scheme of Table 1.2 (N=2).

| Binary | P S T | |
|---|---|---|
| | Positive mode | Negative mode |
| 0 0 | - + | - + |
| 0 1 | 0 + | 0 - |
| 1 0 | + 0 | - 0 |
| 1 1 | + - | + - |

Table 1.2 - Paired Selected Ternary Code

The mode is reversed each time a binary 10 or 01 block is encountered.

The purpose of this code is to eliminate dc and the ternary character 00, simultaneously. In fact, since timing information must be extracted from the pulse train be regenerative repeaters, long sequences of zeros result in long periods without timing information.

We see that there are two possible ternary characters encoding the binary blocks 01 and 10: the dc is balanced by alternating the two kinds of blocks.

Codes with K=4, N=3

With N=3, $3^3$=27 possible ternary blocks are available. A code can be designed by choosing properly the correspondence between binary and ternary blocks.

The basic ideas are:

a. Eliminate the ternary character 000

b. The following ternary blocks: 0 + -    + 0 -    + - 0
0 - +    - 0 +    - + 0

are dc-free and can be used without any restriction.

c. The remaining 20 have non-zero dc: ten of them have positive dc, and ten negative. They can be paired to represent ten different binary blocks; with each binary block, one will associate either one ternary block or the other so as best to compensate the dc.

The first such code is called 4B-3T (Jessop,1968). Here, the characters in each pair are inverses of each other; e.g.:

| | | |
|---|---|---|
| + + + | and | - - - |
| + 0 0 | and | - 0 0 |
| | etc. | |

Another code, described by Franaszek (1968), is based on a more sophisticated attribution of ternary blocks. Frananszek's MS43 code is generated under constant monitoring of the <u>running</u> <u>digital</u> <u>sum</u> (RDS), defined as

$$\sigma_\nu = \sum_{n=\mu}^{\nu} b_n \tag{1.19}$$

where $\mu$ is arbitrary but fixed. When studying a code without a dc component, it is very important to know the maximum <u>digital</u> <u>sum</u> <u>variation</u> (DSV) of this code. The DSV, defined as

$$\mathrm{DSV} = \max_\nu \sigma_\nu - \min_\nu \sigma_\nu \tag{1.20}$$

is a parameter that measures roughly the distortion suffered by a coded signal passing through a channel that does not transmit direct current.

It can be shown that 4B-3T has

DSV = 7

whereas MS43 code has

DSV = 5 .

The coding rule of Franaszek's MS43 is summarized in Table 1.3 .

TABLE 1.3

Franaszek's MS43 code (RDS is computed at the end of the preceding codeword)

| Binary | Ternary | Codeword | |
|---|---|---|---|
| Block | RDS = 1 | RDS = 2,3 | RDS = 4 |
| 0000 | + + + | - + - | |
| 0001 | + + 0 | 0 0 - | |
| 0010 | + 0 + | 0 - 0 | |
| 0100 | 0 + + | - 0 0 | |
| 1000 | + - + | | - - - |
| 0011 | 0 - + | | |
| 0101 | - 0 + | | |
| 1001 | 0 0 + | | - - 0 |
| 1010 | 0 + 0 | | - 0 - |
| 1100 | + 0 0 | | 0 - - |
| 0110 | - + 0 | | |
| 1110 | + - 0 | | |
| 1101 | + 0 - | | |
| 1011 | 0 + - | | |
| 0111 | - + + | | - - + |
| 1111 | + + - | + - - | |

A more sophisticated scheme can be obtained by allowing the codewords to have several different lengths. In Franaszek's VL43 code (Franaszek, 1968), binary data are framed in blocks of 4 or 8 bits, each of them being coded into a ternary block with 3 or 6 symbols, respectively. For this code, the value of DSV is 4.

## B. Non-alphabetic ternary codes

Many non-alphabetic ternary coding schemes have been proposed for use in cable PCM (see Croisier,1970, and references therein).

A scheme which is actually used is the bipolar-with-six-zero-substitution (B6ZS) code (Davis,1969). The basic idea is the following: the source sequence is first encoded using bipolar code, then every sequence of 6 consecutive zeros is substituted by a "filling sequence". Since the filling sequences must be recognized by the receiver, they must contain a bipolar violation; i.e., a pulse whose polarity is the same as that of the last nonzero pulse. In B6ZS, the filling sequence is

$$B0VB0V$$

where B represents a normal bipolar pulse, and V a bipolar violation pulse. As an example, the sequence $\cdots 1010000000110 \cdots$ can be coded as $\cdots +0-\underline{+0+-0-}0+-0\cdots$ , where the filling sequence has been underlined.

## C. Multilevel Codes (M>3)

In digital microwave transmission systems, using either terrestrial radio links or satellite channels, the need of modulation schemes that employ efficiently bandwidth and power has led to the extensive use of phase-shift keying (PSK). In our framework, we can represent M-ary PSK ($M=2^m$) in this way: the binary source sequence is first framed into blocks of $m=\log_2 M$ symbols. To each block, the coder associates the complex number, or "phase",

$$e^{j\ell(2\pi/M)} \qquad \ell=0,1,\ldots,M-1$$

The complex envelope of modulator output is then

$$e^{j\ell(2\pi/M)}\, s(t)$$

where

$$s(t) = \begin{cases} 1 & |t|<T/2 \\ 0 & \text{elsewhere} . \end{cases}$$

PSK works very well under not-too-severe conditions. However, since the power spectrum of a PSK signal exhibits sidelobes that can interfere with neighboring channels, a certain amount of filtering is necessary after the modulator. This filtering results in a great amount of envelope variations in the signal, that pose considerable problems due to the nonlinear elements usually present in a communication system. In fact, limiters, upconverters, traveling-wave tube amplifiers, etc., operated at or near saturation for better efficiency, have AM-PM conversion effects which may seriously impair system performance.

An important fact is that, as I shall show later, the degrading effects of nonlinear devices such as those encountered in practical systems are only envelope-dependent. Thus, modified PSK schemes which exhibit nearly constant envelope after filtering are called for.

It has been experienced that, in conventional PSK, the most critical situation, with respect to envelope fluctuations, occurs when the phase of the transmitted signal incurs a 180° change. In correspondence to such a transition, the signal envelope usually drops to very small values (possibly zero).

To avoid this unwanted effect, several schemes have been devised. The basic idea underlying all such schemes is simple: to avoid transitions between signal points that are symmetric with respect to the origin of signal space. Let us see how modulation and coding schemes can be designed in order to meet this requirement (see also Ajmone and Biglieri,1977).

4-phase offset PSK (See Gronemeyer and McBride,1976, and references therein)

Observing Fig. 1.6, we see that, for a phase transition of 180° to take place, it is necessary that both in-phase and quadrature parts of the signal change sign simultaneously. To avoid this, we can modify a PSK signal by staggering its in-phase and quadrature parts according to the scheme of Fig. 1.6 .

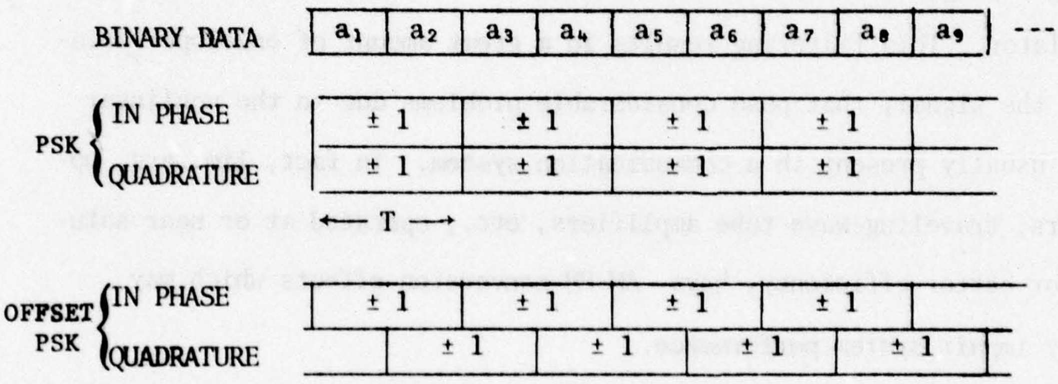


Fig. 1.6

This is actually equivalent to the following (uncoded) modulation scheme:

| binary source pair | modulator waveform |
|---|---|
| 00 | s(t) + js(t-T/2) |
| 01 | s(t) - js(t-T/2) |
| 10 | -s(t) + js(t-T/2) |
| 11 | -s(t) - js(t-T/2) |

where $s(t)=1$ for $|t|<T/2$ and 0 elsewhere.

Similar results can be obtained by coding the source sequence; an example is described below.

Tamm-Danilov codes (Tamm and Danilov,1968)

The basic idea with these codes is to use only a subset of the available symbols for coding; which subset is actually used depends on the source symbol. As an example, assume that six coded symbols are available, namely:

$$e^{j\ell(2\pi/6)} \qquad \ell=0,1,\ldots,5$$

and the coder operates according to the following (differential) rule:

| binary source pair $a_n$ | coded symbol $b_n$ |
|---|---|
| 00 | $b_{n-1}$ |
| 01 | $b_{n-1}e^{j\pi/3}$ |
| 10 | $b_{n-1}e^{2j\pi/3}$ |
| 11 | $b_{n-1}e^{-j\pi/3}$ |

It can be easily seen from the coding rule that 180° transitions cannot take place.

## REFERENCES

M.R. Aaron (1962), "PCM transmission in the exchange plant," BSTJ, vol 41, p. 99 ff, January.

M. Ajmone and E. Biglieri (1977), "Spectral occupancy of complex PSK." Proc. 1977 Intl.Comm.Conf., Chicago, Ill.

A. Croisier (1970), "Introduction to pseudoternary transmission codes," IBM J.Res.Devlop., p.354 ff, July.

J.H. Davis (1969), "T2: a 6.3 Mb/s digital repeatered line," Proc. 1969 Intl. Comm.Conf., Boulder, Colo, p. 34 ff.

P.A. Franaszek (1968), "Sequence-state coding for digital transmission," BSTJ, vol 47, p. 143 ff.

E. Gorog (1968), "Redundant alphabets with desirable frequency spectrum properties," IBM J.Res.Develop., p. 234 ff, May.

S.A. Gronemeyer and A.L. McBride (1976), "MSK and offset QPSK modulation," IEEE Trans. on Commun., vol COM-24, n. 8, p. 809 ff, August.

M. Hecht and A. Guida (1969), "Delay modulation," IEEE Proc, vol. 57, p. 1314 ff, July.

A. Jessop (1968), "High capacity PCM multiplexing and code translation," IEE Electronics Div.Colloq. on PCM, IEE Coll. Digest n. 1968/7, pp. 14/2-14/5, March.

H. Kobayashi (1970), "Coding schemes for reduction of intersymbol interference in data transmission systems," IBM J.Res.Develop., p. 343 ff, July.

H. Kobayashi (1971), "A survey of coding schemes for transmission or recording of digital data," IEEE Trans.Commun.Technol., vol. COM-19, p1087 ff, Dec.

A. Lender (1966), "Correlative level coding for binary-data transmission," IEEE Spectrum, vol. 3, p. 104 ff, February.

W.C. Lindsey and M.K. Simon (1973), Telecommunication System Engineering, (Prentice-Hall, Englewood Cliffs, NJ).

H. Nyquist (1924), "Certain factors affecting telegraph speed," BSTJ, vol 3, p. 324 ff, April.

H. Nyquist (1928), "Certain topics in telegraph transmission theory," AIEE Trans., vol. 47, p. 617 ff, April.

J.M. Sipress (1965), "A new class of selected ternary pulse transmission plans for digital transmission lines," IEEE Trans.Commun.Technol., vol COM-13, p. 366 ff.

Yu. A. Tamm and B.S. Danilov (1968), "Methods of double relative phase-shift keying with partial sidelobe suppression," Telecomm.Radio Eng., vol 22, p. 32 ff.

# CHAPTER 2 - THE DESIGN OF CODING AND MODULATION SCHEMES

## 2.1 TIME-DOMAIN AND FREQUENCY-DOMAIN CONSTRAINTS

As we have seen through several examples presented in Chapter 1, two different philosophies are involved in the design of coding and modulation schemes for channels with memory.

If we look at the discrete channel formed by modulator, transmission medium and demodulator, the goal of the coding scheme will be to restrict the symbol stream to form only "good" sequences; i.e., sequences that are well matched to the channel structure.

If we look instead at the continuous channel, or transmission medium, the goal will be to create the best possible discrete channel. This is obtained by constraining the transmitted signal to match the medium characteristics. A typical approach is to concentrate the transmitted signal power into spectral regions where the channel behavior is better.

In this chapter, we shall see some general techniques useful for the analysis and the design of coding and modulation schemes.

## 2.2 TIME-DOMAIN CONSTRAINTS

The approach taken is based on the use of finite-state machines (see, e.g., Gill,1962) as models of the contraints we want to put on the coded sequence (as an example, we may want to bound the DSV, or the number of consecutive zeros).

Consider, for instance, the constraint that no more than two 0's may be transmitted successively in a binary code. We can represent this constraint using the three-state machine whose state-transition diagram is depicted in Fig. 2.1 .

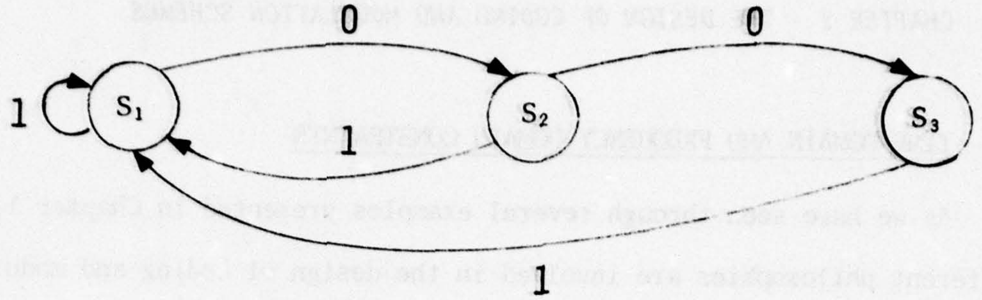


Fig. 2.1

To each source symbol, a transition between states is associated which accounts for the constraints we put on the coded sequence. The mechanism of these transitions is described by a finite-state machine.

We can represent an N-state machine by giving its "skeleton matrix" $\underset{\sim}{D}=(d_{ij})$ ; i.e., an N × N matrix whose entries are given by

$$d_{ij} = \begin{cases} 1 & \text{if transition from state } s_i \text{ to state } s_j \\ & \text{is allowed} \\ 0 & \text{elsewhere} \end{cases} \tag{2.1}$$

and the N × N "symbol-transition matrix" $\underset{\sim}{\Gamma} = (\gamma_{ij})$ , whose entries are

$$\gamma_{ij} = \begin{cases} \text{symbol associated with transition } s_i \rightarrow s_j \\ \qquad \text{if } d_{ij}=1 \\ \emptyset \qquad \text{if } d_{ij}=0 \end{cases} \tag{2.2}$$

In our previous example, we have

$$\underset{\sim}{D} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \tag{2.3}$$

and

$$\underset{\sim}{\Gamma} = \begin{bmatrix} 1 & 0 & \emptyset \\ 1 & \emptyset & 0 \\ 1 & \emptyset & \emptyset \end{bmatrix} \tag{2.4}$$

Clearly, $\underset{\sim}{D}$ is completely determined by $\underset{\sim}{\Gamma}$, so we can say that $\underset{\sim}{\Gamma}$ characterizes the constraints put on the coded sequence. Once these constraints are given, we may ask: how much information can be carried by this sequence? More precisely, we may want to compute the number of bits per symbol carried by the sequence (clearly, the more severe the constraints, the less information will be carried by each symbol).

Define

$m_\nu$ = the number of allowable sequences of length $\nu$;

then the maximum rate of the channel, i.e., the maximum number of bits that can be carried by each coded symbol is (Shannon,1948)

$$C = \lim_{\nu\to\infty} \frac{1}{\nu} \log m_\nu \qquad \text{bits/symbol} \tag{2.5}$$

Consider now the $\nu$-th power of $\underset{\sim}{D}$; if $d_{ij}^{(\nu)}$ denotes the i-j-th entry of $\underset{\sim}{D}^\nu$, and we define

$$A(\nu) = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij}^{(\nu)} \tag{2.6}$$

then upper and lower bounds to $C$ are given by (Franaszek,1969):

$$\frac{1}{N+\nu}\left[\log_2 A(\nu) - 2\log_2 (N+1)\right] \le C \le \frac{1}{\nu} \log_2 A(\nu) \qquad \nu=1,2,\ldots \tag{2.7}$$

The bounds given by (2.7) become tighter as $\nu$ increases, so that $C$

can be approximated within any desired accuracy.

If $R$ denotes the information rate of a code designed to transmit on a channel with maximum rate $C$ , the <u>efficiency</u> of the code can be defined as

$$\eta = \frac{R}{C} \tag{2.8}$$

As an example, upper bounds on the efficiency of codes designed by putting limits on the DSV (see 1.3) can be found in Chien (1970).

Consider now the sequences that can be produced by a finite-state machine characterized by a symbol-transition matrix $\underset{\sim}{\Gamma}$ . The set of sequences of length $\nu$ can be obtained (Franaszek,1970) by taking the $\nu$-th power of the matrix $\underset{\sim}{\Gamma}$ , where ordinary sums and products are substituted by the operations of disjunction (V) and concatenation, respectively. Concatenation of the null symbol $\emptyset$ with any symbol results in $\emptyset$ .

As an example, consider $\nu$=2 and $\underset{\sim}{\Gamma}$ given by (2.4). The possible triplets of symbols are given by

$$\underset{\sim}{\Gamma}^3 = \begin{bmatrix} 111\mathrm{v}011\mathrm{v}101\mathrm{v}001 & 110\mathrm{v}010 & 100 \\ 111\mathrm{v}011\mathrm{v}101 & 110\mathrm{v}010 & 100 \\ 111\mathrm{v}101 & 110 & 100 \end{bmatrix} \tag{2.10}$$

The set of possible $\nu$-tuples is then obtained by taking the disjunction of all the entries of $\underset{\sim}{\Gamma}^{\nu}$ . In our examples, the allowable triplets are all the possible triplets of binary symbols, except 000 .

## Design of alphabetic codes

Suppose the source sequence is partitioned into blocks of length N, each one being mapped by the coder into a block of length N' .

The choice of the codeword used to represent a given source block may be a function of the state to which the finite-state machine representing the coder is taken. In this case, the code is said to be state-dependent. Or the code may be state-independent, which means that codewords can be freely concatenated without violating the constraints.

Consider a simple example; each one of these codewords:

$$\begin{matrix} 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&1&0\\ 1&0&1\\ 0&1&1\\ 1&1&1 \end{matrix} \qquad (2.11)$$

satisfies the constraints of Fig. 2.1 (no more than two 0's may be transmitted successively); but a concatenation of the first and second codeword, for example, would violate them. To get a set of codewords that does not violate the constraint, we must reduce the set (2.11) to 5 elements:

$$\begin{matrix} 1&1&1\\ 1&0&1\\ 0&1&1\\ 1&1&0\\ 0&1&0 \end{matrix} \qquad (2.12)$$

A method for designing an (N,K) alphabet code has been proposed by Franaszek (1968,1969,1970); see also Freiman and Wyner (1964). First, for N and K fixed, a recursive search technique is used (Franaszek, 1969) to determine the existence of a set of principal states. These are states from each of which there exists a number $\beta \geq B^K$ of paths terminating at other principal states. The existence of a set of principal states is a necessary and sufficient condition for the existence of an $(N,K)$ alphabetic code. The words available for encoding are the paths of length $N$ connecting the principal states; these words can be obtained from the $N$-th power of the matrix $\underset{\sim}{\Gamma}$. An example will illustrate this procedure.

<u>Example</u> (Franaszek,1970)

Consider a binary code for a binary source $(B=m=2)$, and

$$\underset{\sim}{D} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad \underset{\sim}{\Gamma} = \begin{bmatrix} \emptyset & 0 & \emptyset & \emptyset \\ 1 & \emptyset & 0 & \emptyset \\ 1 & \emptyset & \emptyset & 0 \\ 1 & \emptyset & \emptyset & \emptyset \end{bmatrix}$$

(see Fig. 2.2). For these constraints, we have $C=.552$ .

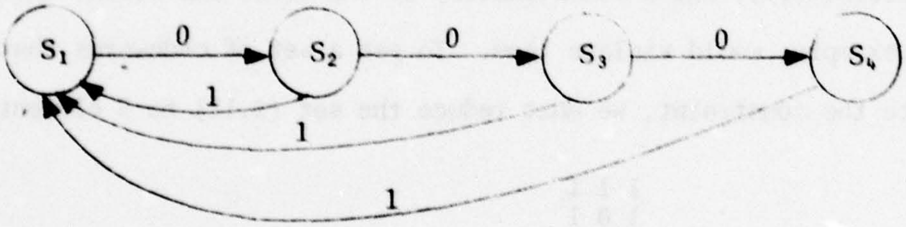


Fig. 2.2

Let $K=1, N=2$ , and look for the existence of a set of principal states. To do this, compute first $\underset{\sim}{D}^N = \underset{\sim}{D}^2$ :

$$\underset{\sim}{D}^2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

It is easily seen that $\{S_1, S_2, S_3\}$ is a set of principal states, from each of which there are $2 = B^K$ paths of length 2 terminating at other principal states. So a (2,1) code exists, with an efficiency of approximately 90% .

Words of length 2 can be obtained from

$$\underset{\sim}{\Gamma}^2 = \begin{bmatrix} 01 & \emptyset & 00 & \emptyset \\ 01 & 10 & \emptyset & 00 \\ 01 & 10 & \emptyset & \emptyset \\ \emptyset & 10 & \emptyset & \emptyset \end{bmatrix}$$

The words available for encoding are the word sets associated with the principal states:

$S_1 \rightarrow$ 01, 00
$S_2 \rightarrow$ 01, 10
$S_3 \rightarrow$ 01, 10

and a code can be constructed according to the following rule:

| binary symbol | state | code-word |
|---|---|---|
| 1 | $S_1$ | 01 |
|  | $S_2$ | 01 |
|  | $S_3$ | 01 |
| 0 | $S_1$ | 00 |
|  | $S_2$ | 10 |
|  | $S_3$ | 10 |

It must be noted that, in general, there are several degrees of freedom in the choice of codewords to be associated to source symbols. This freedom may possibly be exploited to meet some other requirement (such as the minimization of error probability).

## 2.3 A GENERAL TECHNIQUE FOR COMPUTING THE SPECTRUM OF A DIGITAL SIGNAL

In this analysis, I shall assume the following model for the generation of digital signals to be sent through the transmission channel: every T seconds, the modulator emits a waveform chosen in the set

$$\{q(t;k)\}_{k=1}^{M} \tag{2.13}$$

where $q(t;k)$, $1 \le k \le \mu$ , are complex functions of time. Thus, the signal at the output of the modulator is

$$x(t) = \sum_{n=-\infty}^{\infty} q(t-nT;\xi_n) \tag{2.14}$$

where $(\xi_n)_{n=-\infty}^{\infty}$ is a wide-sense stationary sequence of random variables taking values in the set $\{1,2,\dots,M\}$ .

With these assumptions, let us compute the power spectrum of the random process $x(t)$. As shown in Appendix A, we need first to compute the Fourier transform of $x(t)$ , which is given by

$$(\omega) = \sum_{n=-\infty}^{\infty} Q(\omega;\xi_n)e^{-jn\omega T} \tag{2.15}$$

where $Q(\omega;i)$, $1 \le i \le M$, are the Fourier transforms of $q(t;i)$, $1 \le i \le M$ .

The function $\Gamma(\omega_1,\omega_2)$ can be computed according to (A.4), which gives

$$\Gamma(\omega_1,\omega_2) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} E\{Q(\omega_1;\xi_n)Q^*(\omega_2;\xi_m)\}e^{-j(n\omega_1-m\omega_2)T} \tag{2.16}$$

We are assuming that the sequence $(\xi_n)_{n=-\infty}^{\infty}$ is wide-sense stationary; thus, the average in (2.16) will depend only on the difference $n-m$. Defining

$$\mathscr{R}(\omega_1,\omega_2;n-m) \overset{\Delta}{=} E\, Q(\omega_1;\xi_n)Q^*(\omega_2;\xi_m) \tag{2.17}$$

we can rewrite (2.16) as

$$\Gamma(\omega_1,\omega_2) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \mathscr{R}(\omega_1,\omega_2;n-m)e^{-j(n-m)\omega_1 T} e^{-jm(\omega_1-\omega_2)T} \tag{2.18}$$

With the change $\ell=n-m$ in summation indices, and using the equality

$$\sum_{k=-\infty}^{\infty} e^{-jkxy} = \frac{2\pi}{x} \sum_{k=-\infty}^{\infty} \delta(y-k\frac{2\pi}{x}) \tag{2.19}$$

Eq.(2.18) takes the form

$$\Gamma(\omega_1,\omega_2) = \frac{2\pi}{T} \sum_{m=-\infty}^{\infty} \left( \sum_{\ell=-\infty}^{\infty} \mathscr{R}(\omega_1,\omega_2;\ell)e^{-j\ell\omega_1 T}\right)\delta(\omega_1-\omega_2-m\frac{2\pi}{T}) \tag{2.20}$$

Comparing (2.20) to (A.3), we see that the function $\Gamma(\omega_1,\omega_2)$ has only linear masses lying parallel to the bisector of the plane $(\omega_1,\omega_2)$ (actually, the process $x(t)$ is cyclostationary). The power spectrum of $x(t)$ is thus given by

$$\mathscr{G}(\omega) = \frac{1}{T} \sum_{\ell=-\infty}^{\infty} \rho(\omega;\ell)e^{-j\ell\omega T} \tag{2.21}$$

where

$$\rho(\omega;\ell) \overset{\Delta}{=} \mathscr{R}(\omega,\omega;\ell) \; . \tag{2.22}$$

Notice that the series (2.21) may not converge in the usual sense. In fact, we may extract the so-called discrete part of the spectrum, defining

$$\rho(\omega;\infty) \overset{\Delta}{=} \lim_{\ell\to\infty} \mathscr{R}(\omega,\omega;\ell) = |E\{Q(\omega;\xi_n)\}|^2 \tag{2.23}$$

(Of course, we must make the assumption that the limit exists; this condition, as we shall see soon, is met if the sequence $(\xi_n)_{n=-\infty}^{\infty}$ forms a regular Markov chain.) Then, if $\mathscr{G}_c(\omega)$ and $\mathscr{G}_d(\omega)$ represent the continuous and the discrete part (line spectrum) of $\mathscr{G}(\omega)$, respectively, we have, using (2.19) once more:

$$\begin{cases} \mathscr{G}(\omega) = \mathscr{G}_c(\omega) + \mathscr{G}_d(\omega) \\ \mathscr{G}_c(\omega) = \dfrac{1}{T} \displaystyle\sum_{\ell=-\infty}^{\infty} [\rho(\omega;\ell) - \rho(\omega;\infty)]e^{-j\ell\omega T} \\ \mathscr{G}_d(\omega) = \dfrac{2\pi}{T^2} \rho(\omega;\infty) \displaystyle\sum_{\ell=-\infty}^{\infty} \delta(\omega-\ell\dfrac{2\pi}{T}) \end{cases} \tag{2.24}$$

where

$$\rho(\omega;\ell) = E\{Q(\omega;\xi_{\ell+k})Q^*(\omega;\xi_k)\}$$

It can be noted that the line spectrum (which is often an unwanted feature) appears only if $E\{Q(\omega;\xi_n)\} \neq 0$ .

<u>Example</u> (linear modulation)

Assume that the transmitted signal is of the form

$$x(t) = \sum_{n=-\infty}^{\infty} c_n q(t-nT) \tag{2.25}$$

where $E\{c_n\}= 0$ .

Then

$$\rho(\omega;\ell) = E\{c_{\ell+k}\, c_k^*\}|Q(\omega)|^2 = \rho_\ell |Q(\omega)|^2 \tag{2.26}$$

where $(\rho_\ell)_{\ell=-\infty}^{\infty}$ is the covariance sequence of the process $(c_n)_{n=-\infty}^{\infty}$, and $Q(\omega)$ is the Fourier transform of the pulse $q(t)$ ; hence

$$\mathcal{G}(\omega) = \mathcal{G}_c(\omega) = \frac{1}{T}\,|Q(\omega)|^2 \sum_{\ell=-\infty}^{\infty} \rho_\ell\, e^{-j\ell\omega T} \tag{2.27}$$

It can be seen from (2.27) that the spectrum expression is made up of two terms. The first one, $|Q(\omega)|^2$, depends only on the shape of the basic pulse used for modulation. The other term, an infinite summation, depends only on the correlation between coded symbols. Under the assumption that source symbols are independent and identically distributed, this term accounts for the code only.

## 2.4 SOME EXAMPLES OF APPLICATION

The computation of the covariance may not be an easy task; so Eq.(2.27), and more generally Eq.(2.24), may be difficult to apply in the actual computation of power spectra. We shall see later how, with some additional requirements on the sequence $(\xi_n)_{n=-\infty}^{\infty}$, a computationally practical technique can be devised. In this section, we shall restrict our attention to some special cases in which (2.24) can be applied.

Consider first the important special case that arises when the random variables $\xi_n$ are independent and identically distributed. This situation occurs, with our assumptions, when no coding is performed on the source sequence. We shall now see, through an example, how a modulation scheme can be designed on the basis of frequency constraints.

Consider digital transmission over a satellite channel. A sensible criterion in the choice of the modulation scheme turns out to be its spectrum occupancy. In fact, for most planned or implemented satellite links, data streams from users are assigned adjacent frequency bands, or "channels", that interfere with each other to a larger or lesser extent depending on the bandwidth occupancy of the modulated signals.

For PSK modulation, and in general for radio systems, bandwidth occupancy can be defined as the frequency interval that contains a specified fraction of the total radio-frequency power. Since bandwidth is a measure of adjacent channel interference, it should be kept to a minimum.

Consider first conventional quaternary (i.e., M=4) PSK. It fits model (2.14) provided that we define

$$q(t;\xi_k) = e^{j\phi_k} s(t) \tag{2.28}$$

where

$$s(t) = \begin{cases} 1 & |t|<T/2 \\ 0 & \text{elsewhere} \end{cases} \tag{2.29}$$

and

$$\phi_k = (2\xi_k - 1)\pi/4 \tag{2.30}$$

If the random variables $\xi_k$ are independent, identically distributed and take values in $\{1,2,3,4\}$ with equal probabilities, we get from (2.27):

$$\mathcal{G}(\omega) = T\left(\frac{\sin \omega T/2}{\omega T/2}\right)^2 \tag{2.31}$$

Consider now offset PSK , as defined in §1.3 .

We can write

$$q(t;\xi_k) = \varepsilon_k' \, s(t) + j\varepsilon_k'' \, s(t-T/2) \tag{2.32}$$

where $s(t)$ is again given by (2.23), and $\varepsilon_k'$ , $\varepsilon_k''$ are obtained from $\xi_k$ according to the following rule:

| $\xi_k$ | $\varepsilon_k'$ | $\varepsilon_k''$ |
|---|---|---|
| 1 | $1/\sqrt{2}$ | $1/\sqrt{2}$ |
| 2 | $-1/\sqrt{2}$ | $1/\sqrt{2}$ |
| 3 | $-1/\sqrt{2}$ | $-1/\sqrt{2}$ |
| 4 | $1/\sqrt{2}$ | $-1/\sqrt{2}$ |

It turns out that $\varepsilon_k'$ , $\varepsilon_k''$ are independent, identically distributed random variables.

Taking the Fourier transform of $q(t;\xi_k)$ , we get

$$Q(\omega;\xi_k) = T(\varepsilon_k' + j\varepsilon_k'' \, e^{-j\omega T/2}) \, \frac{\sin \omega T/2}{\omega T/2} \tag{2.33}$$

and, consequently,

$$\rho(\omega;\ell) \overset{\Delta}{=} E\{Q(\omega;\xi_{\ell+k})Q^*(\omega;\xi_k)\} \tag{2.34}$$

$$= \begin{cases} T^2\left(\dfrac{\sin \omega T/2}{\omega T/2}\right)^2 & \ell = 0 \\ 0 & \ell \neq 0 \end{cases}$$

so that offset PSK is spectrally equivalent to PSK .

Consider now a more general scheme of offset PSK, that I shall call Shaped Offset PSK. Assume that

$$q(t;\xi_k) = \epsilon_k' f(t) + j\,\epsilon_k'' f(t-T/2) \tag{2.35}$$

where f(t) is a pulse shape of T seconds duration, such that the signal x(t) (see (2.14)) has constant envelope. (This is a feature of PSK signals that we may want to keep.) If $\epsilon_k'$, $\epsilon_k''$ are obtained from $\xi_k$ according to the same rule prescribed for offset PSK, we can easily get, for the spectrum of shaped offset PSK:

$$\mathcal{G}(\omega) = \frac{1}{T}|F(\omega)|^2 \tag{2.36}$$

where $F(\omega)$ is the Fourier transform of f(t).

We would like to find f(t), defined in the interval (-T/2,T/2) and giving a constant envelope x(t), such that the power of x(t) is a maximum in a given frequency band.

Without the constant-envelope restriction, this problem has a classical solution offered by truncated versions of the prolate spheroidal wave functions (see Landau and Pollak, 1961). With this further constraint, the problem is unsolved, so we must restrict ourselves to a less ambitious goal.

Consider the choice

$$f(t) = cos[\frac{\pi}{T} t + \phi(t)] \ , \quad |t|<T/2 \tag{2.37}$$

which leads to a constant envelope x(t) if $\phi(t)$ is periodic with period T/2. Choosing $\phi(t)$ in order to get a sharp roll-off of $\mathcal{G}(\omega)$ as $\omega\to\infty$, we may thus expect that more power is concentrated in the neighborhood of the origin.

Observe that, if $f(t)$ is L-times differentiable, integrating by parts we get, as $\omega \to \infty$:

$$F(\omega) = \sum_{\ell=0}^{L} (-1)^{\ell} \frac{f^{(\ell)}(T/2)e^{-j\omega T/2} - f^{(\ell)}(-T/2)e^{j\omega T/2}}{(j\omega)^{\ell+1}} + 0(\omega^{-L-2}) \tag{2.38}$$

i.e., the behavior of $F(\omega)$, and hence $|F(\omega)|^2$, as $\omega$ grows to infinity, depends on the values of the derivatives of $f(t)$ at the edges of the interval $(-T/2, T/2)$. In particular, if we put

$$f^{(\ell)}(T/2) = f^{(\ell)}(-T/2) = 0 \,, \qquad 0 \le \ell \le L \tag{2.39}$$

then

$$|F(\omega)|^2 = 0(\omega^{-2L-4}) \qquad \omega \to \infty \quad . \tag{2.40}$$

Expand $\phi(t)$ in a Fourier series:

$$\phi(t) = a_1 \cos \frac{4\pi}{T} t + a_2 \cos \frac{4\pi}{T} t + \dots$$
$$+ b_1 \sin \frac{4\pi}{T} t + b_2 \sin \frac{4\pi}{T} t + \dots \quad . \tag{2.41}$$

We may choose the Fourier coefficients in order to achieve a prescribed number of zero derivatives, i.e., a prescribed asymptotic behavior of the spectrum. If all coefficients are zero, i.e.,

$$\phi(t) \equiv 0 \quad , \tag{2.42}$$

we get

$$\Phi(\omega) = 0(\omega^{-4}) \tag{2.43}$$

and this modulation scheme is known as MSK . Using (2.36), the MSK spectrum can be easily computed (see also Simon, 1976):

$$\mathscr{G}(\omega) = \frac{4\pi^2}{T}\left(\frac{\cos\ \omega T/2}{\pi^2-\omega^2T^2}\right)^2 \tag{2.44}$$

Taking $a_1=a_2=\ldots=b_2=b_3=\ldots=0$ and $b_1= -\frac{1}{4}$

we obtain

$$\mathscr{G}(\omega) = 0(\omega^{-6}) \tag{2.45}$$

and a modulation scheme known as SFSK (Amoroso, 1976). Other schemes are presente' by Ajmone and Biglieri (1977) where a comparison between the actual power spectra of PSK , MSK , SFSK and other schemes is reported.

As another application, consider a linear modulator (see (2.25)) driven by a linearly coded sequence $(c_n)_{n=-\infty}^{\infty}$ ; this sequence is obtained from the source sequence $(a_n)_{-\infty}^{\infty}$ according to the rule

$$c_n = \Gamma(D)a_n \tag{2.46}$$

where D is the delay operator, and $\Gamma(\cdot)$ is a polynomial.

If $(a_n)$ is a sequence of zero-mean independent, identically distributed random variables, with a proper scaling of their amplitudes we can assume that its correlation sequence is equal to $(\ldots 0,0,1,0,0,\ldots)$ . In this case, the correlation sequence of $(c_n)$ has a D-transform given by

$$\sum_{\ell=-\infty}^{\infty} \rho_\ell\, D^\ell = \Gamma(D^{-1})\Gamma(D) \tag{2.47}$$

so that, using (2.27):

$$\mathscr{G}(\omega) = \frac{1}{T}|Q(\omega)|^2\ \Gamma(e^{j\omega T})\Gamma(e^{-j\omega T}) \tag{2.48}$$

The term $\Gamma(e^{j\omega T})\Gamma(e^{-j\omega T})$ accounts for the spectrum shaping achieved by coding. Appropriate choice of polynomial $\Gamma(D)$ will lead to the desired spectrum shaping.

For example, the condition $\Gamma(1)=0$ will lead to a zero at the origin. The polynomial $\Gamma(D)=1-D$ (giving rise to twinned binary code, see §1.3) satisfies this condition.

Similarly, the spectrum has a zero at the origin as well as at $\omega=\pi/T$ if $\Gamma(1)=\Gamma(-1) = 0$ . This condition is met by the polynomial $\Gamma(D) = 1-D^2$ (class-IV partial response).

## 2.5 DESIGN OF CODES IN THE FREQUENCY DOMAIN FOR LINEAR MODULATION

Consider the situation represented in the example of §2.3 (linear modulation). Assume that the sequence $(c_n)_{n=-\infty}^{\infty}$ is a real sequence; so we obtain, if the code is linear,

$$\mathcal{G}(\omega) = \frac{1}{T}\,|Q(\omega)|^2\ |\Gamma(e^{j\omega T})|^2 \qquad (2.49)$$

We want to find the coefficients of $\Gamma(\cdot)$ to yield a prescribed spectral shaping

$$\mathcal{G}(\omega) = \frac{1}{T}\,|Q(\omega)|^2 \left\{\sigma_0 + 2\sum_{\ell=1}^{\nu} \sigma_\ell \cos \ell\omega T\right\}$$

$$= \frac{1}{T}\,|Q(\omega)|^2\ C(\omega) \qquad (2.50)$$

The trigonometric polynomial $C(\omega)$ of (2.50) must be non-negative. Otherwise, we would get negative values for $\mathcal{G}(\omega)$, the power spectrum of the coded sequence. As a consequence of a discrete version of the Bochner theorem (the Herglotz lemma: see, e.g., Loève, 1963), since $C(\omega)$ is a non-negative trigonometric polynomial, from its coefficients $(\sigma_\ell)_{\ell=0}^{\nu}$ a finite length autocorrelation sequence $(\rho_\ell)_{\ell=-\infty}^{\infty}$ can be constructed according to the rule

$$\rho_\ell = \begin{cases} \sigma_\ell & \ell=0,1,\dots,\nu \\ \sigma_{-\ell} & \ell=-1,\dots,-\nu \\ 0 & \text{otherwise} \end{cases} \qquad (2.51)$$

(see Gilchrist and Thomas, 1976).

Thus, a polynomial $\Gamma(\cdot)$ such that

$$|\Gamma(e^{j\omega T})|^2 = C(\omega) \tag{2.52}$$

can be found. In fact, due to a theorem on trigonometric polynomials (Szegö,1975), a polynomial $\Gamma(\cdot)$ of degree $\nu$ can always be constructed such that (2.52) holds. This polynomial is generally not unique.

Unfortunately, the coefficients of the coding polynomial $\Gamma(\cdot)$ need not be integer numbers, so that if the source alphabet has $B$ symbols, the coded sequence will take up to $B^{\nu+1}$ different values, which would make the code impractical. Moreover, the problem of characterizing covariance sequences of processes taking only a finite number of values has no solution comparable in simplicity to that provided by the Herglotz lemma when the restriction on allowed values is removed.

As an example, consider the class $\mathcal{U}$ of covariance sequences of processes taking only the values ±1 .

Let $\mathscr{C}_k$ denote the set of all $k \times k$ matrices $\underset{\sim}{A} = (a_{ij})$ with the property that

$$\sum_{i=1}^{k} \sum_{j=1}^{k} x_i \, a_{ij} \, x_j \geq 0 \tag{2.53}$$

where $x_1, \dots, x_k \in \{+1,-1\}$ . Then a sequence $(\rho_n)_{n=-\infty}^{\infty}$ belongs to $\mathcal{U}$ if and only if (McMillan,1955):

(i) $\rho_n$ , $-\infty < n < \infty$ , are real numbers such that $\rho_n = \rho_{-n}$ and $\rho_o = 1$ ;

(ii) for any integer $k > 0$ , any set $\{n_1, n_2, \dots, n_k\}$ of indices and any matrix $\underset{\sim}{A} \in \mathscr{C}_k$ , the following holds

$$\sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} \, \rho_{n_i - n_j} \geq 0 \, .$$

The design problem in the frequence domain can be further generalized if the assumption of linearity of the code is relaxed, while the hypothesis of linear modulation is maintained. Before stating the problem in its generality, let us look at a simple example, which will provide some motivation for the formulation.

Consider once again the twinned binary code. With this ternary code, we get a zero in the spectrum at the origin but a rate of only one information bit per coded symbol, whereas a rate of $log_2 3 \cong 1.58$ bits/symbol could be achieved by a ternary sequence of independent random variables. Thus, a twinned binary code achieves a desirable feature in the spectrum at a cost of about a 1/3 decrease in information rate.

So, it is reasonable to raise the question of whether any scheme with the same spectrum as twinned binary could attain a greater rate. Moreover, what is the greater rate so achievable?

Stated in its most general terms, the problem is the following: Consider a stationary, time discrete random process whose variables $(c_n)_{n=-\infty}^{\infty}$ take their value from a finite set of real numbers. Let the process have mean zero and a given correlation sequence $(\rho_n)_{n=-\infty}^{\infty}$. What is the largest entropy that this process can have, and what is its probability structure?

Such a problem has been considered by Slepian (1972), and left essentially unsolved.

## 2.6 COMPUTATION OF THE POWER SPECTRUM WHEN THE SEQUENCE $(\xi_n)$ FORMS A MARKOV CHAIN

So far, in the derivation of the power spectrum I have made only the assumption that the sequence $(\xi_n)_{n=-\infty}^{\infty}$ is wide-sense stationary. A further specialization of eqs.(2.24) can be obtained through the assumption that such a sequence forms a regular, homogeneous Markov chain. (Notice that, to meet this requirement, it is sometimes necessary to assume that in the set (2.13) some of the waveforms are equal.)

This further assumption on the statistics of the sequence $(\xi_n)$ allows a powerful computational algorithm to be obtained for the derivation of the power spectrum of a modulated signal.

To justify the assumption that the sequence $(\xi_n)$ can be modeled as a Markov chain, it should suffice to observe that, if we model the encoder as a finite-state machine driven by a stationary sequence of independent source symbols, then the sequence of states of the machine forms a homogeneous Markov chain.

The resulting closed-form expression for the spectrum should allow an encoder and a modulator to be designed in order to achieve a given spectral behavior of the transmitted signals. No results, however, are available yet in this area.

The assumption that $(\xi_n)_{n=-\infty}^{\infty}$ is a homogeneous Markov chain means that the one-step transition probabilities

$$\Pr\{\xi_{n+1} = j \,|\, \xi_n = i\} \qquad i,j=1,2,\dots,M \tag{2.54}$$

are independent of the value of n. In this case, we can define the transition probability matrix of the process as the $M \times M$ matrix $\underset{\sim}{P}=(p_{ij})$ with entries

$$p_{ij} \overset{\Delta}{=} \Pr\{\xi_{n+1}= j\,|\,\xi_n= i\} . \tag{2.55}$$

(The i-th row of $\underset{\sim}{P}$ is the probability distribution of $\xi_{n+1}$ , given that $\xi_n=i$ .)

The further assumption that the chain is regular means that the limit

$$\underset{\sim}{P}^{\infty} = \lim_{n\to\infty} \underset{\sim}{P}^{n} \tag{2.56}$$

exists; the matrix $\underset{\sim}{P}^{\infty}$ is idempotent, i.e.,

$$(\underset{\sim}{P}^{\infty})^2 = \underset{\sim}{P}^{\infty} \tag{2.57}$$

and

$$(\underset{\sim}{P}^{\infty})_{ij} = p_j \tag{2.58}$$

where

$$p_j \overset{\Delta}{=} \Pr\{\xi_n= j\} \qquad j=1,\dots,M \tag{2.59}$$

describes the probability distribution of the random variables $\xi_n$ .

The column vector of these probabilities can be obtained by solving the system of linear equations

$$\begin{cases} \underset{\sim}{P}^T\, \underset{\sim}{p} = \underset{\sim}{p} \\ \underset{\sim}{1}^T\, \underset{\sim}{p} = 1 \end{cases} \tag{2.60}$$

where $(\cdot)^T$ denotes transpose, and $\underset{\sim}{1}$ is a column vector all of whose entries are 1 .

Recall now (2.24); defining the column vector

$$\underset{\sim}{Q}(\omega) = [Q(\omega;1),Q(\omega;2),\ldots,Q(\omega;M)]' \tag{2.61}$$

and the diagonal matrix $\underset{\sim}{\Pi}$ whose non-zero entries are

$$(\underset{\sim}{\Pi})_{ii} = p_i \qquad i=1,\ldots,M \tag{2.62}$$

we can write

$$\rho(\omega;\ell) \overset{\Delta}{=} E\{Q(\omega;\xi_{\ell+k})Q^*(\omega;\xi_k)\}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} P(\xi_{\ell+k}= j\,|\,\xi_k= i)P(\xi_k= i)Q(\omega;j)Q^*(\omega;i)$$

$$= \underset{\sim}{Q}^{\dagger}(\omega)\ \underset{\sim}{\Pi}\ \underset{\sim}{P}^{\ell}\ \underset{\sim}{Q}(\omega) \tag{2.63}$$

( $^{\dagger}$ denotes conjugate transpose).

Taking the limit as $\ell\to\infty$ :

$$\rho(\omega;\infty) = \underset{\sim}{Q}^{\dagger}(\omega)\ \underset{\sim}{\Pi}\ \underset{\sim}{P}^{\infty}\ \underset{\sim}{Q}(\omega) \tag{2.64}$$

Thus,

$$\mathcal{S}_C(\omega) = \frac{1}{T}\sum_{\ell=-\infty}^{\infty} [\rho(\omega;\ell) - \rho(\omega;\infty)]e^{-j\ell\omega T} \quad =$$

observing that $\rho(\omega;-n) = \rho^*(\omega;n)$ ,

$$= \frac{2}{T}\,\mathrm{Re}\left\{\sum_{\ell=0}^{\infty} [\rho(\omega;\ell)- \rho(\omega;\infty)]e^{-j\ell\omega T}\right\} - \frac{1}{T}\,[\rho(\omega;0)- \rho(\omega;\infty)] =$$

using (2.63)-(2.64),

$$= \frac{2}{T}\,\mathrm{Re}\left\{\underset{\sim}{Q}^{\dagger}(\omega)\,\underset{\sim}{\Pi}\sum_{\ell=-\infty}^{\infty} (\underset{\sim}{P}^{\ell}-\underset{\sim}{P}^{\infty})e^{-j\ell\omega T}\ \underset{\sim}{Q}(\omega)\right\} - \frac{1}{T}\,\underset{\sim}{Q}^{\dagger}(\omega)\underset{\sim}{\Pi}(\underset{\sim}{I}-\underset{\sim}{P}^{\infty})\underset{\sim}{Q}(\omega) \tag{2.65}$$

Observe now that $\underset{\sim}{P}^{\infty}\underset{\sim}{P} = \underset{\sim}{P}\underset{\sim}{P}^{\infty} = \underset{\sim}{P}^{\infty}$ ; thus

$$\underset{\sim}{P}^{\ell}-\underset{\sim}{P}^{\infty} = \begin{cases} \underset{\sim}{I}-\underset{\sim}{P}^{\infty} & \ell=0 \\ (\underset{\sim}{P}-\underset{\sim}{P}^{\infty})^{\ell} & \ell\neq 0 \end{cases} \tag{2.66}$$

So we can rewrite (2.65) as follows:

$$\mathscr{G}_c(\omega) = \frac{2}{T}\,\mathrm{I\!Re}\left\{\underset{\sim}{Q}^{\dagger}(\omega)\underset{\sim}{\Pi}\underset{\sim}{\Lambda}(\omega)\underset{\sim}{Q}(\omega)\right\} - \frac{1}{T}\,\underset{\sim}{Q}^{\dagger}(\omega)\underset{\sim}{\Pi}(\underset{\sim}{I}+\underset{\sim}{P}^{\infty})\underset{\sim}{Q}(\omega) \tag{2.67}$$

where

$$\underset{\sim}{\Lambda}(\omega) = \sum_{\ell=0}^{\infty} (\underset{\sim}{P}-\underset{\sim}{P}^{\infty})^{\ell}\, e^{-j\ell\omega T} \ . \tag{2.68}$$

In Appendix B, some techniques are shown for the computation of the matrix series (2.68).

In a similar way, we can derive the expression for the discrete part of the spectrum:

$$\mathscr{G}_d(\omega) = \frac{2\pi}{T^2}\ \underset{\sim}{Q}^{\dagger}(\omega)\underset{\sim}{\Pi}\ \underset{\sim}{P}^{\infty}\underset{\sim}{Q}(\omega) \sum_{n=-\infty}^{\infty} \delta(\omega-n\,\frac{2\pi}{T}) \tag{2.69}$$

Eqs. (2.67) and (2.69) were first obtained, in scalar form, by Tausworthe and Welch (1961). Subsequently, they were independently rediscovered by several other researchers.

For a further specialization of (2.67)-(2.69) to constant-length alphabetic codes, see Cariolaro and Tronca (1974). Gilchrist and Thomas (1975) consider the computation of the power spectrum for algebraic error-correcting codes.

No result seems to be available on the power spectrum shaping obtained using convolutional codes; the analysis could be carried out as a rather straightforward application of (2.67) ÷ (2.69).

An example

Consider the Miller code (§1.2) . The coded sequence can be represented as a 4-state Markov chain with transition probability matrix

$$\underset{\sim}{P} = \frac{1}{2}\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

The vector $\underset{\sim}{Q}(\omega)$ is given by

$$\underset{\sim}{Q}(\omega) = G(\omega)\begin{bmatrix} 1+e^{-j\omega T/2} \\ 1-e^{-j\omega T/2} \\ -1+e^{-j\omega T/2} \\ -1-e^{-j\omega T/2} \end{bmatrix}$$

where $G(\omega)$ is the Fourier transform of

$$g(t) = \begin{cases} 1 & 0\le t<T/2 \\ 0 & \text{elsewhere} \end{cases} .$$

We get

$$\underset{\sim}{P}^{\infty} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

and

$$\underset{\sim}{A} \overset{\Delta}{=} \underset{\sim}{P} - \underset{\sim}{P}^{\infty} = \frac{1}{4}\begin{bmatrix} -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \end{bmatrix}$$

By application of the Faddeev algorithm (Appendix B), we get

$$\delta_1 = 1 \; ; \quad \delta_2 = \frac{1}{2} \; ; \quad \delta_3 = \delta_4 = 0$$

$$\underset{\sim}{B}_1 = \frac{1}{4}\begin{bmatrix} 3 & 1 & -1 & 1 \\ -1 & 3 & 1 & 1 \\ 1 & 1 & 3 & -1 \\ 1 & -1 & 1 & 3 \end{bmatrix}$$

$$\underset{\sim}{B}_2 = \frac{1}{4}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$$\underset{\sim}{B}_3 = 0 \; .$$

In conclusion, we get

$$\mathscr{G}(\omega) = \mathscr{G}_C(\omega) = \frac{T}{2}\left(\frac{\sin \omega T/4}{\omega T/4}\right)^2 \cdot \frac{3 + \cos \omega T/2 + 2\cos \omega T - \cos 3\omega T/2}{9 + 12\cos \omega T + 4\cos 2\omega T}$$

## REFERENCES

M. Ajmone and E. Biglieri (1977), "Spectral occupancy of complex PSK", Proc. 1977 Intl.Comm.Conf., Chicago, Ill.

M. Ajmone and E. Biglieri (1977), "Power spectra of complex PSK for satellite communications", to appear in Alta Frequenza.

F. Amoroso (1976), "Pulse and spectrum manipulation in the minimum (frequency) shift keying (MSK) format", IEEE Trans.Commun.Technol., vol. COM-17, p. 581 ff, March.

G.L. Cariolaro and G.P. Tronca (1974), "Spectra of block coded digital signals", IEEE Trans on Commun., vol. COM-22, n.10, p. 1555 ff, October.

Ta-Mu Chien (1970), "Upper bound on the efficiency of dc-constrained codes", BSTJ, vol. 49, p. 2267 ff, November.

P.A. Franaszek (1968), "Sequence-state coding for digital transmission", BSTJ, vol. 47, p. 143 ff.

P.A. Franaszek (1969), "On synchronous variable length coding for discrete noiseless channels", Information and Control, vol. 15, p. 155 ff, August.

P.A. Franaszek (1970), "Sequence-state methods for run-length-limited coding", IBM J.Res.Develop., p. 376 ff, July.

C.V. Freiman and A.D. Wyner (1964), "Optimum block codes for noiseless input restricted channels", Information and Control, vol. 7, p. 398 ff.

J.H. Gilchrist and J.B. Thomas (1975), "Power spectral density of modulated error-correcting coded sequences", IEEE Trans. on Commun., vol. COM-23, n.11, p. 1207 ff, November.

J.H. Gilchrist and J.B. Thomas (1976), "Synthesis of spectral shaping block codes for PAM", IEEE Trans. on Inform.Theory, vol. IT-22, n.5, p. 546 ff, September.

A. Gill (1962), Introduction to the Theory of Finite-State Machines, McGraw-Hill, New York (publ).

H.J. Landau and H.O. Pollack (1961), "Prolate spheroidal wave functions, Fourier analysis and uncertainty-II", BSTJ, vol. 40, p. 65 ff, January.

M. Loève (1963), Probability Theory (van Nostrand Co., Princeton, publ.)

B. McMillan (1955), "History of a problem", SIAM Journal, vol. 3, n.3, p. 119 ff, September.

C.E. Shannon (1948), "A mathematical theory of communication", BSTJ, vol.27, p. 379 ff, and 623 ff.

M.K. Simon (1976), "A generalization of MSK-type signalling based upon input data symbol pulse shaping", IEEE Trans. on Commun., vol. COM-24, p. 845 ff, August.

D. Slepian (1972), "On maxentropic discrete stationary processes", BSTJ, vol. 51, n.3, p. 629 ff, March.

G. Szegö (1975), Orthogonal polynomials , in Am.Math.Soc.Coll.Publ., vol vol. XXIII, pp. 2-5.

R.C. Tausworthe and L.R. Welch (1961), "Power spectra of signals modulated by random and pseudorandom sequences", JPL Tech.Rept., No. 32-140, Pasadena, CA.

# CHAPTER 3 - NONLINEAR CHANNELS WITH MEMORY

## 3.1 INTRODUCTION

In this chapter, I shall consider the problem of characterizing nonlinear channels with memory. Actually, analysis of these channels, as well as evaluation of the performance of digital transmission schemes, is an important practical problem. For example, in telephone lines used for data transmission, the advent of equalization, and hence of new precision in transmission, revealed that nonlinear distortion -- arising principally from inaccuracy in companding -- is a serious source of performance impairment. It has been conjectured that, for data transmission systems operating at rates greater than 4800 bits/s , the error rate is almost entirely determined by nonlinear distortion (Lucky,1975).

Another important example of nonlinear channel with memory arises from digital satellite communications. The on-board amplifiers, operated at or near saturation for better efficiency, exhibit strongly nonlinear characteristics.

The Volterra-series approach has been taken because it provides the most general analytical tool to deal with nonlinear channels with memory. Although it suffers some drawbacks -- for example, it cannot be parametrized, not unlike the impulse response for characterizing a linear channel -- its generality makes it attractive in several instances.

## 3.2 CHARACTERIZING A NONLINEAR CHANNEL

For linear channels, the input-output relationship is fully described by their impulse response. Similarly, if the channel is nonlinear but memoryless, and sufficiently well-behaved, an input-output relationship can be obtained by expanding the nonlinear characteristics in a power series. More generally, for a nonlinear channel with memory that satisfies certain regularity conditions, a generalization of these output representations is provided by Volterra series. Entry to the literature on this subject can be made through Bedrosian and Rice (1971). Here I shall present, mostly in a heuristic way, some of the basic features of this theory and certain of its applications.

To provide some motivation for the general input-output relationship of a nonlinear system with memory, let us consider a simple example.

Assume that the system is created (Fig. 3.1) by cascading a linear, time-invariant system with impulse response h(t) and a nonlinear, memoryless system with an analytic input-output relationship $z(t)=f[v(t)]$ .

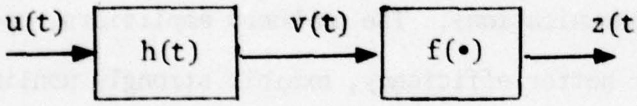


FIG. 3.1

Let

$$f(\cdot) = \sum_{k=1}^{\infty} \gamma_k \frac{(\cdot)^k}{k!} \qquad (3.1)$$

be the Taylor series expansion of $f(\cdot)$ (assume that $f(0)=0$ , so the term with $k=0$ is missing).

Denoting by $z(t)$ the output when the input is $u(t)$, we easily get, under suitable regularity assumptions for the functions involved:

$$\begin{aligned} z(t) &= f\left[\int_{-\infty}^{\infty} h(\tau)u(t-\tau)d\tau\right] = \\ &= \sum_{k=1}^{\infty} \frac{\gamma_k}{k!}\left[\int_{-\infty}^{\infty} h(\tau)u(t-\tau)d\tau\right]^k = \\ &= \sum_{k=1}^{\infty} \frac{\gamma_k}{k!} \int_{-\infty}^{\infty} d\tau_1 \int_{-\infty}^{\infty} d\tau_2 \cdots \int_{-\infty}^{\infty} d\tau_k \cdot \prod_{r=1}^{k} h(\tau_r) \cdot \prod_{r=1}^{k} u(t-\tau_r) \end{aligned} \tag{3.2}$$

Letting

$$h_k(\tau_1,\tau_2,\ldots,\tau_k) \overset{\Delta}{=} \frac{\gamma_k}{k!} \prod_{r=1}^{k} h(\tau_r) \tag{3.3}$$

the input-output relationship takes the following form:

$$z(t) = \sum_{k=1}^{\infty} \int_{-\infty}^{\infty} d\tau_1 \int_{-\infty}^{\infty} d\tau_2 \cdots \int_{-\infty}^{\infty} d\tau_k h_k(\tau_1,\tau_2,\ldots,\tau_k) \cdot \prod_{r=1}^{k} u(t-\tau_r) \tag{3.4}$$

Eq.(3.4), without assumption (3.3), is the most general form of input-output relationship of a time-invariant nonlinear system with memory that meets certain regularity conditions. It can be seen that the "Volterra kernels" $h_k(\tau_1,\tau_2,\ldots,\tau_k)$, a generalization of impulse response for linear systems, completely describe the system behavior. Thus, the problem of characterizing a nonlinear system with memory reduces to the problem of computing its Volterra kernels.

We have seen that, for the simple channel obtained by cascading a linear system and a nonlinear channel with memory, the Volterra kernels are given by Eq.(3.3). In a similar way, it can be shown that, for a system which can be modeled as a memoryless nonlinearity $f(\cdot)$ preceded and followed by two linear systems (Fig. 3.2),

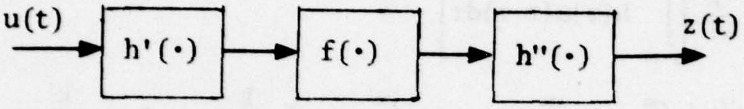


FIG. 3.2

we get

$$h_k(\tau_1,\ldots,\tau_k) = \frac{\gamma_k}{k!}\int_{-\infty}^{\infty} d\tau \; h''(\tau) \prod_{r=1}^{k} h'(\tau_r-\tau) \tag{3.5}$$

where $\gamma_k$ are the power series coefficients of $f(\cdot)$, and $h'(\cdot), h''(\cdot)$ are the impulse responses of the linear systems preceding and following the nonlinearity, respectively.

Eq.(3.5) has a simple interpretation, which may also be useful to compute numerically the Volterra kernels. Fix $\tau_1,\ldots,\tau_{s-1},\tau_{s+1},\ldots,\tau_k$; then rewrite (3.5) as follows:

$$h_k(\tau_1,\ldots,\tau_k) = \frac{\gamma_k}{k!}\int_{-\infty}^{\infty} d\tau \left[ h''(\tau) \prod_{\substack{r=1 \\ r\neq s}}^{k} h'(\tau_r-\tau) \right] h'(\tau_s-\tau)$$

This expression, apart from a constant coefficient, can be viewed as the response, at time $\tau_s$ , of a linear system with impulse response $h'(\cdot)$ to the input

$$h''(t) \prod_{\substack{r=1 \\ r\neq s}}^{k} h'(\tau_r - t) \quad .$$

Using standard numerical filtering techniques, if $\gamma_k$, $h'(\cdot)$ and $h''(\cdot)$ are known, it is possible to compute any kernel for any prescribed set of time instants $\tau_1,\ldots,\tau_k$ .

It can be observed that, in both equations (3.3) and (3.5), the kernels are symmetrical functions of their arguments; i.e., kernel values do not change if their arguments are permuted. It can be shown (Bedrosian and Rice,1971) that the assumption of symmetry for the kernels does not entail any loss of generality so, throughout this chapter, all Volterra kernels will be assumed to be symmetric, unless otherwise specified.

## 3.3 SOME EXAMPLES

In this section, the expression of the output from a channel described by a Volterra series will be specialized to some cases of practical relevance. In particular, a digital signal, a general harmonizable random process and a sum of two such processes will be considered as inputs to the channel. The corresponding outputs will be derived in a form that will allow some analysis of their statistics.

Consider first a baseband, linearly modulated digital signal

$$x(t) = \sum_{k=-\infty}^{\infty} c_k \, q(t-kT) \tag{3.6}$$

where $(c_n)_{n=-\infty}^{\infty}$ is a real, discrete-time stationary random process.

We can assume that (3.6) is obtained by passing the following signal:

$$u(t) = \sum_{k=-\infty}^{\infty} c_k \, \delta(t-kT) \tag{3.7}$$

through a linear, time-invariant system whose impulse response is $q(t)$. This linear system will be included in the channel structure, and the Volterra kernels of the channel will be modified accordingly.

If the signal (3.7) is sent through a channel described by the general input-output relationship (3.4), we get at the output

$$v(t) = \sum_{k=1}^{\infty} \sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} c_{n_1} c_{n_2} \cdots c_{n_k} \, h_k(t-n_1T, t-n_2T, \dots, t-n_kT) \tag{3.8}$$

where the indices $n_1, n_2, \dots, n_k$ run from $-\infty$ to $\infty$. In particular, if we sample the signal $v(t)$ at $t=t_0$, we get

$$v(t_0) = \sum_{k=1}^{\infty} \sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} c_{n_1} c_{n_2} \cdots c_{n_k} \, H_k(n_1, n_2, \dots, n_k) \tag{3.9}$$

where

$$H_k(n_1, n_2, \dots, n_k) \overset{\Delta}{=} h_k(t_0-n_1T, t_0-n_2T, \dots, t_0-n_kT) \ . \tag{3.10}$$

Eq.(3.9) can also be rewritten as follows:

$$v(t_0) = c_0 \, H_1(0) + \sum_{n_1 \neq 0} c_{n_1} H_1(n_1) + \sum_{k=2}^{\infty} \sum_{n_1} \cdots \sum_{n_k} c_{n_1} \cdots c_{n_k} \cdot H_k(n_1, \dots, n_k) \tag{3.11}$$

where one can easily recognize the various contributions to the signal at the output of the channel: useful signal, intersymbol interference produced by linear distortion, nonlinear interference.

Assume now that the input of the nonlinear channel is a real, harmonizable process $u(t)$. Using the series expansion for these random processes devised by Cambanis and Liu(1971) (a heuristic proof of this result is presented in Appendix C for the special case of wide-sense stationary processes), we can represent $u(t)$ as

$$u(t) = \sum_n \xi_n \, s_n(t) \tag{3.12}$$

where $(\xi_n)_{n=-\infty}^{\infty}$ is a real, discrete-time random process such that

$$E\xi_n\xi_m = \delta_{mn} \quad . \tag{3.13}$$

At the output of the channel, we get

$$z(t) = \sum_{k=1}^{\infty} \sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} \xi_{n_1}\xi_{n_2}\cdots\xi_{n_k} \, \sigma_k(t;n_1,\ldots,n_k) \tag{3.14}$$

where

$$\sigma_k(t;n_1,n_2,\ldots,n_k) = \int_{-\infty}^{\infty} d\tau_1 \int_{-\infty}^{\infty} d\tau_2 \cdots \int_{-\infty}^{\infty} d\tau_k \, h_k(\tau_1,\tau_2,\ldots,\tau_k) \prod_{r=1}^{k} s_{n_r}(t-\tau_r) \tag{3.15}$$

An important special case arises when $s_n(t)$ satisfies the following relation:

$$s_n(t) = s(t-n\Theta) \quad . \tag{3.16}$$

(Besides the special case of digital signals, (3.16) holds true for example when u(t) is a bandlimited white noise: see Appendix C.) In this case

$$s_{n_r}(t-\tau_r) = s(t-\tau_r-n_r\Theta)$$

so that, for each value of k , only one integral has to be computed to obtain the kernels (3.15).

Consider finally the sum of two random processes, say

$$u(t) = x(t) + y(t) \quad .$$

Assume that they can be expanded in the following form

$$x(t) = \sum_n \xi_n \, s_n(t)$$

$$y(t) = \sum_n \eta_n \, r_n(t) \qquad (3.17)$$

Generally, this can be obtained by using the Cambanis-Liu expansion, or using eq.(3.6), if x(t) or y(t) or both are digital signals.

The k-th order term in the Volterra-series expansion of the output of the nonlinearity is given by

$$\int_{-\infty}^{\infty} d\tau_1 \ldots \int_{-\infty}^{\infty} d\tau_k \; h_k(t-\tau_1,\ldots,t-\tau_k) \; .$$

$$\cdot \left[ \sum_{n_1} \left\{ \xi_{n_1} \, s_{n_1}(\tau_1) + \eta_{n_1} \, r_{n_1}(\tau_1) \right\} \right] \cdot$$

$$\cdots$$

$$\cdot \left[ \sum_{n_k} \left\{ \xi_{n_k} \, s_{n_k}(\tau_k) + \eta_{n_k} \, r_{n_k}(\tau_k) \right\} \right] \qquad (3.18)$$

Then, defining

$$\rho_{k-i,i}(t;n_1,\ldots,n_k) = \binom{k}{i}\int_{-\infty}^{\infty} d\tau_1 \cdots \int_{-\infty}^{\infty} d\tau_k \; h_k(t-\tau_1,\ldots,t-\tau_k) \cdot$$

$$\cdot \; s_{n_1}(\tau_1)\cdots s_{n_{k-i}}(\tau_{k-i}) \; r_{n_{k-i+1}}(\tau_{k-i+1})\cdots r_{n_k}(\tau_k) \qquad (3.19)$$

we can write the output of the nonlinearity, say $v(t)$, as

$$v(t) = \sum_{k=1}^{\infty} \sum_{i=0}^{k} \sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} \xi_{n_1} \cdots \xi_{n_{k-i}} \; \eta_{n_{k-i+1}} \cdots \eta_{n_k} \rho_{k-i,i}(t;n_1,\ldots,n_k) \qquad (3.20)$$

Notice that the terms $\rho_{k-i,i}(\cdots)$ account for the interaction between the two processes. In particular, $\rho_{k,0}(\cdots)$, $1 \le k < \infty$, give the output corresponding to the input $x(t)$ alone, whereas $\rho_{0,k}(\cdots)$, $1 \le k < \infty$, do the same thing for the input $y(t)$ alone.

## 3.4 FIRST-ORDER STATISTICS OF THE OUTPUT OF A NONLINEAR CHANNEL

In this section, we shall consider the problem of evaluating the first-order statistics of the signal at the output of a nonlinear channel.

In Appendix D, it is shown how it is possible, from the knowledge of a few moments of the random variable $X$, to derive bounds to quantities like

$$E\{\Omega(X)\}$$

where $\Omega(\cdot)$ is a known function, or

$$\Pr\{X \le \lambda\}$$

where $\lambda$ is a known quantity.

Thus, the problem of deriving the first-order statistics of a random variable can be reduced to the problem of computing its moments. Consider, in particular, the problem of computing the moments of the process at the output of a nonlinear channel with memory. Let us assign a a time instant $t_0$, and write the channel output at $t_0$ as

$$\Xi = \sum_{k=1}^{\infty} \sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} a_{n_1} a_{n_2} \cdots a_{n_k} S_k(n_1, n_2, \ldots, n_k) \quad . \tag{3.21}$$

It can be observed that expression (3.21) encompasses both cases (3.9) and (3.14), with proper definition of the quantities involved.

The problem of computing the moments $E\{\Xi^\ell\}$ will be solved in two steps: first, I shall show that $\Xi^2, \Xi^3, \ldots$ can be given the form of

Volterra series. Second, I shall develop a procedure for averaging a Volterra series (see Benedetto, Biglieri and Daffara (1976)).

Consider the two Volterra series

$$\sum_{k=1}^{\infty} \sum_{n_1} \cdots \sum_{n_k} a_{n_1} \cdots a_{n_k} F_k(n_1,\ldots,n_k) \tag{3.22}$$

and

$$\sum_{k=1}^{\infty} \sum_{n_1} \cdots \sum_{n_k} a_{n_1} \cdots a_{n_k} G_k(n_1,\ldots,n_k) \quad . \tag{3.23}$$

The product between them can be given the form

$$\sum_{k=1}^{\infty} \sum_{n_1} \cdots \sum_{n_k} a_{n_1} \cdots a_{n_k} Q_k(n_1,\ldots,n_k) \tag{3.24}$$

where the resulting kernels $Q_k(\cdots)$ are obtained through the recursion

$$\begin{cases} Q_k(n_1,\ldots,n_k) = \sum\limits_{i=1}^{k-1} F_i(n_1,\ldots,n_i) G_{k-i}(n_{i+1},\ldots,n_k) & k\geq 2 \\ Q_1(n_1) \equiv 0 \end{cases} \tag{3.25}$$

Using this result, we can see that the power $\Xi^{\ell}$ can be given the form

$$\Xi^{\ell} = \sum_{k=\ell}^{\infty} \sum_{n_1} \cdots \sum_{n_k} a_{n_1} \cdots a_{n_k} S_k^{(\ell)}(n_1,\ldots,n_k) \tag{3.26}$$

where the Volterra coefficients $S_k^{(\ell)}(\cdots)$ are obtained recursively as

$$S_k^{(\ell)}(n_1,\dots,n_k) = \begin{cases} \sum_{i=\ell-1}^{k-1} S_i^{(\ell-1)}(n_1,\dots,n_i) S_{k-i}^{(1)}(n_{i+1},\dots,n_k) & k \geq \ell \\ 0 & k < \ell \end{cases} \tag{3.27}$$

with the starting values

$$S_i^{(1)}(n_1,\dots,n_i) = S_i(n_1,\dots,n_i) \tag{3.28}$$

In conclusion, any power of $\Xi$ can be expressed as a Volterra series whose coefficients can be derived through a recurrence relationship. Thus, the computation of the moments of $\Xi$ is reduced to the computation of the average of a Volterra series.

I shall describe in the following a simple procedure for computing such an average, under the hypothesis that the random variables $a_{n_i}$ are statistically independent and equally distributed.

Due to the structure of (3.26), it can be seen that $E\Xi^{\ell}$ will be obtained provided that we are able to compute the average

$$E\{a_{n_1} \dots a_{n_k}\} \quad .$$

Denoting by $\lambda_i$ the number of indices $n_1,\dots,n_k$ taking value i (clearly $\sum_i \lambda_i = k$) , we have

$$E\{a_{n_1} \dots a_{n_k}\} = \prod_i E\{a^{\lambda_i}\} \quad .$$

A similar result applies to the computation of the moments in a situation like that represented by eq.(3.20) (see Benedetto, Biglieri and Daffara (1976) for details), when the sum of two independent processes enters the nonlinear channel.

As an application of these procedures, see Benedetto, Biglieri and Daffara (1976), where the error probability for PAM transmission over a nonlinear channel with memory is computed.

## 3.5 BANDPASS NONLINEAR SYSTEMS

In this Section, the results obtained previously -- an input-output relationship valid for nonlinear systems with memory -- will be specialized to the case of bandpass nonlinear systems.

Consider such a system , and a bandpass input. The analytic signal associated with the input can be expressed as:

$$x(t) = A(t)\, e^{j[\omega_0 t + \theta(t)]} \tag{3.30}$$

where $A(t)$ and $\theta(t)$ are baseband signals, and $\omega_0$ is the center frequency of the power spectrum of $x(t)$ . Letting

$$\psi(t) \overset{\Delta}{=} \omega_0 t + \theta(t) \tag{3.31}$$

we can write (3.30) as

$$x(t) = A(t)\, e^{j\psi(t)} \quad . \tag{3.32}$$

Consider now a nonlinear, memoryless system with input-output relationship given by

$$\begin{aligned} y(t) &= \mathscr{P}[x(t)] \\ &= \mathscr{P}[A e^{j\psi}] \end{aligned} \tag{3.33}$$

where $y(t)$ is the output signal.

This is a periodic function of $\psi$ , with period $2\pi$ . So we can represent it as a Fourier series:

$$y(t) = \sum_{n=-\infty}^{\infty} c_n(A)e^{jn\psi} \tag{3.34}$$

where

$$c_n(A) = \frac{1}{2\pi}\int_0^{2\pi} \mathscr{P}[Ae^{j\psi}]e^{-jn\psi}\, d\psi \tag{3.35}$$

Due to the definition of $\psi$ , (3.34) shows that the power spectrum of $y(t)$ will generally include several spectral zones, centered around multiples of the frequency $\omega_0$ . Suppose then that the output of the nonlinearity is followed by a zonal filter, whose function is to stop all the spectral components other than that centered at $\omega_0$ .

The analytic signal associated with this output component is given by

$$y_1(t) = c_1[A(t)]e^{j\psi(t)} \tag{3.36}$$

where

$$c_1(A) = \frac{1}{2\pi}\int_0^{2\pi} \mathscr{P}[Ae^{j\psi}]e^{-j\psi}\, d\psi\ . \tag{3.37}$$

Since in general $c_1(A)$ is a complex number, we can write it in the following form:

$$c_1(A) = F(A)e^{j\phi(A)} \tag{3.38}$$

so that the output $y_1(t)$ becomes

$$y_1(t) = F[A(t)]e^{j\{\psi(t) + \phi[A(t)]\}} . \tag{3.39}$$

Comparing (3.39) with (3.32), we can see that the effect of a nonlinear bandpass system will be to alter the amplitude as well as the phase of the input according to a law that depends only on $A(t)$ , the envelope of the input.

Thus, to describe a nonlinear bandpass system without memory, it is sufficient to assign two functions $F(A),\phi(A)$ , describing the so-called AM/AM conversion and AM/PM conversion of the system. These functions can be suitably parametrized in order to get a useful characterization of the system in terms of a small set of parameters (see Lindsey et al.,(1977), and references therein).

Consider now the more general case in which the nonlinear system has memory. Rewrite first the input of the system as

$$x(t) = \tilde{x}(t)e^{j\omega_0 t} \tag{3.40}$$

where $\tilde{x}(t)$ , the complex envelope of the input, is defined as

$$\tilde{x}(t) \overset{\Delta}{=} x(t)e^{-j\omega_0 t} . \tag{3.41}$$

The output of the system, under suitable regularity hypotheses, can be written in terms of a Volterra series as

$$y(t) = \sum_{\ell=1}^{\infty} \int_{-\infty}^{\infty} d\tau_1 \dots \int_{-\infty}^{\infty} d\tau_\ell \; h_\ell(\tau_1,\dots,\tau_\ell) \prod_{r=1}^{\ell} \mathrm{Re}[x(t-\tau_r)] \tag{3.42}$$

Now,

$$\mathbb{R}\mathrm{e}[x(t)] = \frac{1}{2}\,[x(t) + x^*(t)]$$

$$= \frac{1}{2}\,[\tilde{x}(t)e^{j\omega_0 t} + \tilde{x}^*(t)e^{-j\omega_0 t}] \qquad (3.43)$$

Observe now that, with $\mathbb{R}\mathrm{e}[x(t)]$ expressed as in (3.43), if we expand the products at RHS of (3.42), each will give rise to a sum of $2^{\ell}$ products. Each of these products involves k factors of the type $exp\{\pm j\omega_0(t-\tau_r)\}$ .

Suppose, as we did before, that only the first-zone spectral components of the signal at the output of the nonlinearity are of interest. This means that among these products we need retain only those which give rise to a factor $exp \pm j\omega_0 t$ .

Thus, we are constrained to consider only the products corresponding to odd values of $\ell$ , say $\ell=2k+1$, $k=0,1,\ldots$ . Among these, we retain the products with k (respectively, k+1) factors of the type $exp\{-j\omega_0(t-\tau_r)\}$ and k+1 (respectively, k) factors of the type $exp\{+j\omega_0(t-\tau_r)\}$ .

There are $\binom{2k+1}{k}$ of these products, and each will give the same value for the $\ell$-fold integral under the assumption of symmetric kernels. Thus, denoting by $y_1(t)$ the first-zone filtered component of the output signal,

$$y_1(t) = \frac{1}{2}\,[e^{j\omega_0 t}\,\tilde{y}(t) + e^{-j\omega_0 t}\,\tilde{y}^*(t)] \ . \qquad (3.43)$$

where

$$\tilde{y}(t) \overset{\Delta}{=} \sum_{k=0}^{\infty} \frac{\binom{2k+1}{k}}{2^{2k}} \int_{-\infty}^{\infty} d\tau_1 \cdots \int_{-\infty}^{\infty} d\tau_{2k+1} h_{2k+1}(\tau_1,\ldots,\tau_{2k+1}) \ \cdot$$

$$\cdot \prod_{r=1}^{k} \tilde{x}^*(t-\tau_r)e^{+j\omega_0\tau_r} \cdot \prod_{s=k+1}^{2k+1} \tilde{x}(t-\tau_s)e^{-j\omega_0\tau_s} \qquad (3.44)$$

Since (3.43) can be rewritten as

$$y_1(t) = \mathrm{Re}[\tilde{y}(t)e^{j\omega_0 t}] \ , \tag{3.45}$$

we can identify $\tilde{y}(t)$ with the complex envelope of $y_1(t)$ .

The expression for $\tilde{y}(t)$ will now be modified in order to define equivalent, low-pass Volterra kernels. For this purpose, we must exploit the hypothesis that the nonlinear device is bandpass. This means that the Fourier transforms of its Volterra kernels:

$$\mathscr{H}_{2k+1}(\omega_1,\omega_2,\ldots,\omega_{2k+1}) \overset{\Delta}{=} \int_{-\infty}^{\infty} d\tau_1 \ldots \int_{-\infty}^{\infty} d\tau_{2k+1} h_{2k+1}(\tau_1,\ldots,\tau_{2k+1}) \cdot$$

$$\cdot\ exp{-j}(\omega_1\tau_1 + \omega_2\tau_2 + \ldots + \omega_{2k+1}\tau_{2k+1}) \tag{3.46}$$

differ significantly from zero only over small neighborhoods of the $2^{2k+1}$ points with coordinates $(\pm\omega_0,\pm\omega_0,\ldots,\pm\omega_0)$ .

Thus, we can write $\mathscr{H}_{2k+1}(\ldots)$ as a sum of $2^{2k+1}$ functions with arguments $(\omega_1 \pm\omega_0, \omega_2 \pm\omega_0,\ldots,\omega_{2k+1} \pm\omega_0)$ . Each of these functions is significantly different from zero only in the neighborhood of the origin.

To prove this, consider, as system input, a sinusoidal signal with frequency $\omega'$ and amplitude $A$ . The complex envelope of the output is, from (3.44):

$$\tilde{y}(t) = \sum_{k=0}^{\infty} \frac{\binom{2k+1}{k}}{2^{2k}} A^{2k+1} \mathscr{H}_{2k+1}(\underbrace{\omega',\omega',\ldots,\omega'}_{k+1},\underbrace{-\omega',\ldots,-\omega'}_{k}) \tag{3.47}$$

For the system to be bandpass, $\tilde{y}(t)$ must be significantly different from zero only for $\omega' \cong \omega_0$. This is actually the case only if the Volterra kernels have a functional structure like that just described.

Take now the inverse Fourier transform of $\mathscr{H}_{2k+1}(\cdots)$ . In general, denoting by $\mathscr{F}_\ell$ the $\ell$-dimensional Fourier-transform operator, we have

$$\mathscr{F}_\ell^{-1}[\mathscr{A}(\omega_1 \pm\omega_0, \omega_2 \pm\omega_0, \ldots, \omega_\ell \pm\omega_0)] = exp\{-j\omega_0(\pm t_1 \pm t_2 \pm \ldots \pm t_\ell)\} \cdot$$

$$\cdot\ \mathscr{F}_\ell^{-1}[\mathscr{A}(\omega_1, \omega_2, \ldots, \omega_\ell)] \quad . \tag{3.48}$$

Thus, the inverse Fourier transform of $\mathscr{H}_{2k+1}(\cdots)$ will be a sum of $2^{2k+1}$ baseband functions (kernels), each one being multiplied by a factor

$$exp\{-j\omega_0(\pm\tau_1 \pm\tau_2 \pm \cdots \pm\tau_{2k+1})\} \quad . \tag{3.49}$$

Observe now that, when such an inverse Fourier transform is substituted for $h_{2k+1}(\cdots)$ in (3.44), the exponential factors in the integral may combine with that appearing in the RHS of (3.48), giving rise to terms $exp \pm jn\omega_0\tau_i$ , $n\neq 0$ . These terms will make the integrand oscillate at frequencies not less than $\omega_0$ , so the corresponding values of the integrals will be relatively small.

Thus, the only term to be retained in the inverse Fourier transform of $\mathscr{H}_{2k+1}(\cdots)$ is the one whose exponential factor cancels out with those appearing in the RHS of (3.44). Thus, for our purposes we can approximate $h_{2k+1}(\tau_1, \ldots, \tau_{2k+1})$ in (3.44) as follows:

$$h_{2k+1}(\tau_1, \ldots, \tau_{2k+1}) = \frac{1}{2} e^{-j\omega_0(\tau_1 + \ldots + \tau_k - \tau_{k+1} - \cdots - \tau_{2k+1})} \tilde{h}_{2k+1}(\tau_1, \ldots, \tau_{2k+1}) \tag{3.50}$$

and call $\tilde{h}_{2k+1}(\cdots)$ , defined in this way, the "equivalent low-pass Volterra kernels". Thus, eq.(3.44) becomes

$$\tilde{y}(t) = \sum_{k=0}^{\infty} \frac{\binom{2k+1}{k}}{2^{2k+1}} \int_{-\infty}^{\infty} d\tau_1 \cdots \int_{-\infty}^{\infty} d\tau_{2k+1}\, \tilde{h}_{2k+1}(\tau_1,\ldots,\tau_{2k+1}) \cdot$$
$$\cdot \prod_{r=1}^{k} \tilde{x}^*(t-\tau_r) \prod_{s=k+1}^{2k+1} \tilde{x}(t-\tau_s) \qquad (3.51)$$

Eq. (3.51) gives the complex envelope of the output of a nonlinear system with memory in terms of the complex envelope of the input and the equivalent low-pass Volterra kernels.

Note in particular that, if we stop the expansion at RHS of (3.51) at the first term (linear system), we get the known result:

$$\tilde{y}(t) = \frac{1}{2} \int_{-\infty}^{\infty} \tilde{h}(\tau)\tilde{x}(t-\tau)d\tau \qquad (3.52)$$

where $\tilde{h}(\cdot)$ is the equivalent low-pass impulse response of the linear system.

<u>Example 1</u>

As a simple example, let us consider a sinusoidal signal input with complex envelope

$$\tilde{x}(t) = A\, e^{j\theta} \qquad (3.53)$$

(this corresponds to the real signal $x(t) = A\, cos(\omega_0 t+\theta)$ ). Then the complex envelope of the output signal is given by

$$\tilde{y}(t) = e^{j\theta} \sum_{k=0}^{\infty} \frac{\binom{2k+1}{k}}{2^{2k+1}} A^{2k+1} \quad \beta_{2k+1} \qquad (3.54)$$

where

$$\beta_{2k+1} = \int_{-\infty}^{\infty} d\tau_1 \ldots \int_{-\infty}^{\infty} d\tau_{2k+1}\, \tilde{h}_{2k+1}(\tau_1,\ldots,\tau_{2k+1}) \tag{3.55}$$

Using now (3.50), we can get

$$\beta_{2k+1} = 2\,\mathcal{H}_{2k+1}(\underbrace{\omega_0,\ldots,\omega_0}_{\longleftarrow k+1 \longrightarrow},\underbrace{-\omega_0,\ldots,-\omega_0}_{\longleftarrow k \longrightarrow}) \tag{3.56}$$

in accordance with the result of Bedrosian and Rice (1971).

<u>Example 2</u>

Consider now the transmission of a bandpass digital signal over this channel. I shall assume that the channel has been modeled in such a way that the complex envelope of the input signal takes the form:

$$\tilde{x}(t) = \sum_{n=-\infty}^{\infty} c_n\, \delta(t-nT) \tag{3.57}$$

where $(c_n)_{n=-\infty}^{\infty}$ is a discrete-time, complex random process.

Using (3.51), we get

$$\tilde{y}(t) = \sum_{k=0}^{\infty} l_k \sum_{n_1} \cdots \sum_{n_{2k+1}} c_{n_1} \ldots c_{n_{k+1}} c^*_{n_{k+2}} \ldots c^*_{n_{2k+1}} \cdot$$

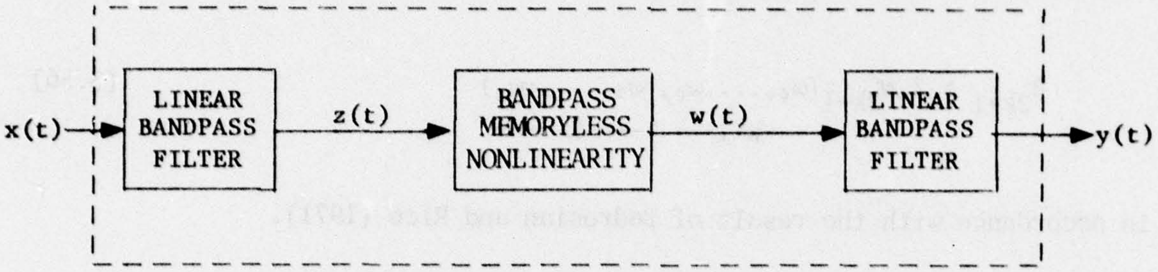
$$\cdot\, \tilde{h}_{2k+1}(t-n_1T,\ldots,t-n_{2k+1}T) \tag{3.58}$$

where

$$l_k = \binom{2k+1}{k} 2^{-2k-1} \quad . \tag{3.59}$$

## 3.6 MODELING A SATELLITE CHANNEL USING BANDPASS VOLTERRA SERIES

As an example of actual computation of a Volterra series, let us consider the channel model represented in Fig. 3.3, which is usually assumed for satellite communications. Here a nonlinear memoryless part, representing the on-board traveling-wave tube, is preceded and followed by two bandpass linear systems. The first represents the cascade of earth station transmitting filter and satellite input filter; the other represents the cascade of satellite output filter and earth station receiving filter.



BANDPASS NONLINEAR SYSTEM

FIG. 3.3

Assume that $h'(\cdot),h''(\cdot)$ are the impulse responses of the two filters, both of whose transfer functions $H'(\cdot),H''(\cdot)$ are centered around the frequency $\omega_0$ ; assume also that the memoryless nonlinear device has an analytic input-output relationship of the form (3.1).

Thus, the Volterra series representation of such a system has kernels given by (3.5). Consider now the complex envelope of the output.

Computing the Fourier transform of the kernels (3.5), we get

$$\mathscr{H}_{2k+1}(\omega_1,\omega_2,\cdots,\omega_{2k+1}) = \frac{\gamma_{2k+1}}{(2k+1)!} H''(\omega_1+\omega_2+\cdots+\omega_{2k+1}) \prod_{r=1}^{2k+1} H'(\omega_r) \tag{3.60}$$

Introduce now the equivalent low-pass transfer functions of the two filters, writing

$$\begin{aligned} H'(\omega) &= \tilde{H}'(\omega-\omega_0) + \tilde{H}'^{*}(-\omega-\omega_0) \\ H''(\omega) &= \tilde{H}''(\omega-\omega_0) + \tilde{H}''^{*}(-\omega-\omega_0) \end{aligned} \tag{3.61}$$

Inserting (3.61) into (3.60), we get

$$\begin{aligned} \mathscr{H}_{2k+1}(\omega_1,\cdots,\omega_{2k+1}) = \frac{\gamma_{2k+1}}{(2k+1)!} [\tilde{H}''(\omega_1+\cdots+\omega_{2k+1}-\omega_0) + \\ + \tilde{H}''^{*}(-\omega_1-\cdots-\omega_{2k+1}-\omega_0)] \prod_{r=1}^{2k+1} \{\tilde{H}'(\omega_r-\omega_0) + \tilde{H}'^{*}(-\omega_r-\omega_0)\} \end{aligned} \tag{3.62}$$

If the products are computed, we obtain a sum of $2^{2k+2}$ terms, only one of which will give rise to a factor $exp\{-j\omega_0(\tau_1+\cdots+\tau_k-\tau_{k+1}-\cdots-\tau_{2k+1})$ after the inverse Fourier transform is taken. This term is

$$\frac{\gamma_{2k+1}}{(2k+1)!} [\tilde{H}''(\omega_1+\cdots+\omega_{2k+1}-\omega_0)] \prod_{r=1}^{k} \tilde{H}'^{*}(-\omega_r-\omega_0) \prod_{s=k+1}^{2k+1} \tilde{H}'(\omega_s-\omega_0) \tag{3.63}$$

as can be seen by computing its inverse Fourier transform, which is

$$exp\{-j\omega_0(\tau_1+\cdots+\tau_k-\tau_{k+1}-\cdots-\tau_{2k+1})\}\cdot$$

$$\cdot\ \frac{\gamma_{2k+1}}{(2k+1)!}\int_{-\infty}^{\infty}\tilde{h}''(\tau)\prod_{r=1}^{k}\tilde{h}'^{*}(\tau_r-\tau)\prod_{s=k+1}^{2k+1}\tilde{h}'(\tau_s-\tau)d\tau$$

where $\tilde{h}'(\cdot),\tilde{h}''(\cdot)$ are the equivalent low-pass impulse responses of the linear filters.

So, using (3.50), we finally obtain

$$\tilde{h}_{2k+1}(\tau_1,\ldots,\tau_{2k+1}) = \frac{\gamma_{2k+1}}{(2k+1)!}\int_{-\infty}^{\infty}\tilde{h}''(\tau)\prod_{r=1}^{k}\tilde{h}'^{*}(\tau_r-\tau)\prod_{s=k+1}^{2k+1}\tilde{h}'(\tau_s-\tau)d\tau \qquad (3.64)$$

It can be observed that

(i) if $H'(\cdot)$ and $H''(\cdot)$, the transfer functions of the two filters, are symmetric around the center frequency $\omega_0$, then the impulse responses $\tilde{h}'(\cdot)$ and $\tilde{h}''(\cdot)$ are real functions, and so is the integral (3.64), and

(ii) if the Volterra kernels are not real, they are not symmetric functions of their arguments; only if their first $k$, or last $k+1$, arguments are permuted is the value of the kernels unchanged.

Let us now turn our attention to the memoryless nonlinear device of Fig. 3.3 . In order to describe this device through an input-output relationship involving complex envelopes, let us observe that the bandpass Volterra series expansion for it can be obtained simply by letting $\tilde{h}'(\cdot) = \tilde{h}''(\cdot) = \delta(\cdot)$ in (3.64) and using (3.51):

$$\tilde{y}(t) = \sum_{k=0}^{\infty}\frac{\gamma_{2k+1}}{k!(k+1)!2^{2k+1}}[\tilde{x}^{*}(t)]^{k}[\tilde{x}(t)]^{k+1} \qquad (3.65)$$

(Notice that only the odd-order terms of the series expansion (3.1) are involved in (3.65). The even-order terms give rise to spectral zones of the output signal in which we are not interested.)

Letting

$$\tilde{x}(t) = A(t)e^{j\theta(t)} \tag{3.66}$$

we can rewrite (3.65) as

$$\tilde{y}(t) = e^{j\theta(t)} \sum_{k=0}^{\infty} \frac{\gamma_{2k+1}}{k!(k+1)!2^{2k+1}} A^{2k+1}(t) \tag{3.67}$$

Since in general the output of a bandpass nonlinear device can be represented as

$$\tilde{y}(t) = F(A)e^{j[\theta(t)+\phi(A)]} \quad , \tag{3.68}$$

(see (3.39)), comparing (3.68) and (3.67) we get

$$F(A)e^{j\phi(A)} = \sum_{k=0}^{\infty} \frac{\gamma_{2k+1}}{k!(k+1)!2^{2k+1}} A^{2k+1} \tag{3.69}$$

For example, if the LHS is represented using a power series:

$$F(A)e^{j\phi(A)} = \sum_{k=0}^{\infty} \frac{f_k}{k!} A^k \tag{3.70}$$

we easily get

$$\gamma_{2k+1} = \frac{2^{2k+1}}{\binom{2k+1}{k}} f_{2k+1} \tag{3.71}$$

Similarly, the LHS of (3.70) can be represented as

$$F(A)e^{j\phi(A)} = \sum_{\ell=1}^{L} b_\ell J_1(\delta_\ell \alpha A) \tag{3.72}$$

where $\alpha$ is a real constant, $b_1,\ldots,b_L$ are complex numbers, and the $\delta_\ell$'s can take the value $\delta_\ell=\ell$ (see, e.g., Shimbo, 1976) or are zeros of $J_1(x)$ (Lindsey *et al*, 1977). In this case, recalling

$$J_1(x) = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{m!(m+1)!2^{2m+1}}$$

we get the simple result

$$\gamma_{2k+1} = \alpha^{2k+1} \sum_{\ell=1}^{L} b_\ell \delta_\ell^{2k+1} \quad . \tag{3.73}$$

## REFERENCES

E. Bedrosian and S.O. Rice (1971). "The output properties of Volterra systems (nonlinear systems with memory) driven by harmonic and Gaussian inputs", IEEE. Proc., vol. 59, p. 1699 ff, December.

S. Benedetto, E. Biglieri and R. Daffara (1976). "Performance of multilevel baseband digital systems in a nonlinear environment", IEEE. Trans. on Commun., vol. COM-24, p. 1166 ff, October.

W.C. Lindsey, J.K. Omura, K.T. Woo, T.C. Huang, L. Biederman (1977). "Investigation of modulation/coding tradeoff for military satellite communications. Vol 2: System modeling and analysis", LINCOM Technical Report, January.

R.W. Lucky (1975). 'Modulation and detection for data transmission on the telephone channel", in J.K. Skwirzynski, ed., New Directions in Signal Processing in Communication and Control, Noordhoff,Leiden (Holland).

O. Shimbo and P.J. Pontano (1976). "A general theory for intelligible crosstalk between frequency-division multiplexed angle-modulated carriers", IEEE Trans. on Commun., vol. COM-24, p. 999 ff, September.

# 4. ANALYSIS OF DIGITAL COMMUNICATION SYSTEMS OPERATING ON NONLINEAR CHANNELS

## 4.1 INTRODUCTION

In this chapter, some of the results obtained before will be used to evaluate the performance of a digital communication system operating on a nonlinear channel with memory.

First, a Markov chain model will be derived for the output of a nonlinear channel whose input is a digital signal. Then, this model will be used to evaluate the power spectrum of such an output, and to derive the structure of the optimum (maximum-likelihood) receiver.

Finally, a Markov chain model will be derived for the discrete channel created by this communication situation.

## 4.2 A MARKOV CHAIN MODEL FOR THE CHANNEL OUTPUT

Assume first, for simplicity's sake, that the nonlinear channel can be modeled as in Chapter 3, and consider the channel output when a linearly modulated digital signal is sent at its input. This output is given by (3.8) or (3.58), according to whether the channel is baseband or bandpass.

Let us assume that the memory of the system is finite. In the Volterra series model, this is equivalent to assuming that all the Volterra kernels that describe the channel are zero -- or reasonably close to zero -- when at least one of their arguments , say $\tau_1$ , takes values outside the interval $(\theta_1,\theta_2)$. (Notice that $\theta_1,\theta_2$ are not dependent on index i , due to the symmetry of the kernels.)

Under this assumption, at any given instant t the output of the channel will depend only on a finite number, say L , of symbols $c_n$ . In fact, all summations with indices $n_i$ in (3.8) and (3.58) will involve a finite number of terms. Thus, we can say that the channel output $v(t)$ is some function of the type

$$v(t) = V(t, c_{\ell_1(t)}, \ldots, c_{\ell_L(t)}) \tag{4.1}$$

where $\ell_1(t), \ldots, \ell_L(t)$ are integer numbers dependent on the value of $t$.

Furthermore, we can say that, if we observe the channel output $v(t)$ for $T$ seconds, this waveform will take on only a finite number of possible shapes. In fact, as $t$ ranges into any finite interval, the integers $\ell_1(t), \ldots, \ell_L(t)$ will take different values, but still in some finite range.

In general, we can say that, if the symbol sequence $(c_n)_{n=-\infty}^{\infty}$ is a stationary random process, observing $v(t)$ for $T$ seconds will give rise to a finite number, say $M'$, of different waveforms. If $L$ denotes the number of values taken by the random variables $c_n$, and $\mathscr{L}$ is the number of symbols on which $v(t)$ depends as $t$ runs in an interval of length $T$, we will get

$$M' \leq L^{\mathscr{L}} \tag{4.2}$$

(the possible inequality accounts for the situation in which the sequence $(c_n)$ is coded).

In conclusion, at the output of the channel, and assuming for the moment that there is no noise, we shall get a situation similar to that occurring at the output of the modulator in the channel model analyzed in

Chapter 2. Every $T$ seconds we shall get a waveform chosen in some finite set $\{q(t;i)\}_{i=1}^{M'}$ , $t \in (0,T)$ , so that we can write

$$v(t) = \sum_{n=-\infty}^{\infty} q(t-nT;\xi_n) \qquad (4.3)$$

where $(\xi_n)_{n=-\infty}^{\infty}$ is a sequence of random variables taking values in the set $\{1,2,\ldots,M'\}$ .

We can analyze this communication situation provided that we are able to determine the statistics of the discrete-time random process $(\xi_n)_{n=-\infty}^{\infty}$ . This is accomplished by assuming, for notationaly simplicity, that as $t$ runs in the interval $[kT,(k+1)T]$ , $v(t)$ depends on $c_k,\ldots,c_{k+\mathscr{L}-1}$ . Therefore, $\xi_k$ will be a function of the same random variables. We can also assume, without loss of generality, that $\xi_k$ is a one-to-one function of these random variables.

Consider now the time interval $[(k+1)T,(k+2)T]$ . Here $v(t)$, and hence $\xi_{k+1}$ , will depend on $c_{k+1},\ldots,c_{k+\mathscr{L}}$ , and so on for other time intervals. The following conclusion can be drawn immediately. Consider $\Pr\{\xi_{k+1}|\xi_k,\xi_{k-1},\ldots\}$ : $\xi_{k+1}$ is a one-to-one function of $c_{k+1},\ldots,c_{k+\mathscr{L}}$, whereas $\xi_k,\xi_{k-1},\ldots$, (i.e., the past of $\xi_{k+1}$) is a one-to-one function of $c_{k+\mathscr{L}-1},c_{k+\mathscr{L}-2},\ldots$ .

Since $(c_k)_{k=-\infty}^{\infty}$ is a Markov chain, the values taken by $c_{k+1},\ldots,c_{k+\mathscr{L}}$ will depend only on $c_k,\ldots,c_{k+\mathscr{L}-1}$ . This means that the values taken by $\xi_{k+1}$ depend only on the value taken by $\xi_k$ , and not on those taken by $\xi_{k-1},\xi_{k-2},\ldots$ . In other words, the sequence $(\xi_k)_{k=-\infty}^{\infty}$ forms a Markov chain.

To compute the transition probability matrix of this chain, assume that $\xi_k$ takes value $i$ when $c_k=i_o$ , $c_{k+1}=i_1,\ldots,c_{k+\mathscr{L}-1}= i_{\mathscr{L}-1}$ .

Similarly, $\xi_{k+1}$ takes value $j$ when $c_{k+1}=j_1,\ldots,c_{k+\mathscr{L}}=j_{\mathscr{L}}$ , and so on. We get

$$\Pr\{\xi_{k+1}=j\,|\,\xi_k=i\} =$$

$$= \Pr\{c_{k+1}=j_1,\ldots,c_{k+\mathscr{L}}=j_{\mathscr{L}}\,|\,c_k=i_0, c_{k+1}=i_1,\ldots,c_{k+\mathscr{L}-1}=i_{\mathscr{L}-1}\}$$

$$=\begin{cases}\Pr\{c_{k+\mathscr{L}}=j_{\mathscr{L}}\,|\,c_{k+\mathscr{L}-1}=i_{\mathscr{L}-1}\} & \text{if } \quad j_\ell=i_\ell\ , \quad 1\le \ell \le \mathscr{L}-1 \\ 0 & \text{elsewhere}\end{cases} \tag{4.4}$$

Eq.(4.4) also shows that, provided that the Markov chain $(c_n)_{n=-\infty}^{\infty}$ is homogeneous, so is the Markov chain $(\xi_n)_{n=-\infty}^{\infty}$ .

Example

Assume $\mathscr{L}=3$ , and a binary signal entering the channel. Under the further assumption that this signal has not been encoded, the random variables $c_n$ turn out to be independent and equally likely, so that the process $(c_n)_{n=-\infty}^{\infty}$ is described by the trivial transition probability matrix:

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

The transition probability matrix of $(\xi_n)_{n=-\infty}^{\infty}$ is $8\times 8$ . Labeling its rows and columns by the corresponding values of the triplets $c_1c_2c_3$ , we get

$$\underline{P} = \frac{1}{2}\begin{array}{cc} \left[\begin{array}{cccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right] & \begin{array}{c} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \\ \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{array} & \end{array}$$

## 4.3 COMPUTATION OF THE POWER SPECTRUM

Using the Markov chain model derived in §4.2, several results about the digital communication system at hand can be obtained in a conceptually straightforward way.

One of these results is the computation of the power spectrum of the channel output. Another, the derivation of the optimum receiver, is discussed in the next section.

Recalling the results described in §2.6, we can see that the spectrum of the digital signal v(t) , the output of a nonlinear channel with memory, can be computed using eqs.(2.67)-(2.69) (Ajmone, Biglieri and Elia,1978). The only hindrance to the application of this method is created by the large dimensionality of the transition probability matrix of the Markov chain $(\xi_n)$; this can seriously impair the usefulness of this technique when the memory of the channel is considerably long.

In any case, several shortcuts can simplify this procedure in order to make it practical. The simplest, and most obvious, is to take advantage of any linear system that can possibly be present at the end of the channel.

Suppose that the nonlinear channel can be split into the cascade of a nonlinear subchannel with memory and a linear system with transfer function $H(\omega)$. This is the case of any practical communication system, where the demodulator is preceded by the receiver filter. Under these conditions, we can compute the power spectrum before this linear system, and then multiply it by $|H(\omega)|^2$.

Another shortcut tries to take advantage of the structure of the matrix $\underset{\sim}{P}$. Consider for example the problem of computing $\underset{\sim}{\Lambda}(\omega)$ as defined in eq.(2.68). This can be done (see Appendix B) by using either Faddeev's algorithm or the closed-form solution (B.6)-(B.10) when the polynomial of the matrix $\underset{\sim}{P}$ is known. The latter technique is more useful in this case. In fact, it can be proved (see Ajmone, Biglieri and Elia,1978) that the minimum polynomial of $\underset{\sim}{P}$ is simply given by $\lambda^{k-1}\Gamma(\lambda)$, $\Gamma(\lambda)$ being the minimum polynomial of the transition probability matrix of the chain $(c_n)$. This fact allows a closed-form solution to be obtained for $\underset{\sim}{\Lambda}(\omega)$ involving the parameters which depend only on the statistics of $(c_n)$.

Similarly, the solution of system (2.6) is generally required to get $\underset{\sim}{\Pi}$ and $\underset{\sim}{P}^{\infty}$ (see (2.58) and (2.62)). This can become difficult as the dimension of the matrix $\underset{\sim}{P}$ increases. Actually, this computation can be reduced to a simpler one by taking advantage of the relation (4.4).

## 4.4 OPTIMUM (MAXIMUM LIKELIHOOD) RECEIVER

We shall now derive the structure of the maximum likelihood receiver for a situation in which the channel consists of a known finite memory part followed by a noisy memoryless part. In particular, we can assume that the received signal is the sum of a signal like (4.3) plus white Gaussian noise. This situation has been considered in a general, abstract setting by Omura (1971), and for the specific problem of a bandpass nonlinear channel by Mesiya et al. (1977).

Since the channel is assumed to have a finite memory, the signal received at time t can be written as

$$z(t) = \sum_{n=n_1(t)}^{n_2(t)} q(t-nT;\xi_n) + \nu(t) \tag{4.5}$$

where $n_1(t)$ and $n_2(t)$, $n_2(t) \geq n_1(t)$ , are integers dependent upon the actual value of t .

Assume now that $z(t)$ is observed over the time period $0 \leq t \leq KT$ , say. Denoting by $N_1$ and $N_2$ the following integers:

$$N_1 = \min_{0 \leq t \leq KT} n_1(t)$$

$$N_2 = \max_{0 \leq t \leq KT} n_2(t) \tag{4.6}$$

we see that the observation will depend on the values taken by the random variables $\xi_{N_1},\ldots,\xi_{N_2}$ . This sequence of random variables may take one of

$$\mathscr{M} = (M')^{N_2-N_1+1} \tag{4.7}$$

AD-A057 375 CALIFORNIA UNIV LOS ANGELES DEPT OF SYSTEM SCIENCE F/G 17/2
CODING AND MODULATION FOR COMMUNICATION CHANNELS WITH MEMORY.(U)
MAR 78 E BIGLIERI N00014-76-C-0875
UNCLASSIFIED UCLA-ENG-7829 NL

2 OF 2
AD
A057375

END
DATE
FILMED
9-78
DDC

possible states, each state corresponding to a received waveform like

$$v_{\underset{\sim}{m}}(t) = \sum_{n=N_1}^{N_2} q(t-nT;m_n) \qquad 0 \le t \le KT \tag{4.8}$$

where $\underset{\sim}{m} = (m_{N_1},\dots,m_{N_2})$ is an integer sequence denoting a possible state taken by the sequence of random variables $\xi_{N_1},\dots,\xi_{N_2}$.

Compute now the log-likelihood ratio for $\underset{\sim}{m}$ ; we get

$$\Lambda_{\underset{\sim}{m}} = \frac{2}{N_0}\int_0^{KT} v_{\underset{\sim}{m}}(t)z(t)dt - \frac{1}{N_0}\int_0^{KT} v_{\underset{\sim}{m}}^{\,2}(t)dt\ . \tag{4.9}$$

Using (4.5), it follows that

$$\Lambda_{\underset{\sim}{m}} = \frac{2}{N_0}\sum_{n=N_1}^{N_2}\int_0^{KT} q(t-nT;m_n)z(t)dt\ -$$

$$-\frac{1}{N_0}\sum_{\ell=N_1}^{N_2}\sum_{n=N_1}^{N_2}\int_0^{KT} q(t-\ell T;m_\ell)q(t-nT;m_n)dt \tag{4.10}$$

Notice now that, under our hypotheses, q(.;.) has a finite duration T. Assuming that K is large enough so that we can disregard end effects, we have

$$\int_0^{KT} q(t-nT;m_n)z(t)dt = \int_{nT}^{(n+1)T} q(t-nT;m_n)z(t)dt \tag{4.11}$$

Similarly we can observe that, owing to the finite duration of q(.;.) ,

$$\int_0^{KT} q(t-\ell T;m_\ell)q(t-nT;m_n)dt = \begin{cases} 0 & \ell\neq n \\ \int_0^{T} q^2(t;m_n)dt & \ell=n \end{cases} \tag{4.12}$$

Thus, defining

$$\alpha_n(m_n) \overset{\Delta}{=} \int_{nT}^{(n+1)T} q(t;m_n)z(t)dt \tag{4.13}$$

and

$$\mathscr{E}(m_n) \overset{\Delta}{=} \int_{0}^{T} q^2(t;m_n)dt \tag{4.14}$$

we finally get

$$\Lambda_{\underset{\sim}{m}} = \frac{2}{N_0} \sum_{n=N_1}^{N_2} \alpha_n(m_n) - \frac{1}{N_0} \sum_{n=N_1}^{N_2} \mathscr{E}(m_n)$$

$$= \frac{1}{N_0} \sum_{n=N_1}^{N_2} \{2\alpha_n(m_n) - \mathscr{E}(m_n)\} . \tag{4.15}$$

We can observe that:

(i) $\alpha_n(m_n)$ can be obtained as the output, sampled at time $(n+1)T$, of a filter matched to $q(t;m_n)$ to the input $z(t)$, $nT \le t < (n+1)T$ .

(ii) $\mathscr{E}(m_n)$ is the energy of the waveform $q(t;m_n)$ .

The maximum likelihood sequence decoding rule now requires $\Lambda_{\underset{\sim}{m}}$ to be maximized over the set of $\mathscr{M}$ possible sequences $\underset{\sim}{m}$ . Observe now the structure of $\Lambda_{\underset{\sim}{m}}$ given by eq.(4.16) which, with obvious notations, we can rewrite as

$$\Lambda(\underset{\sim}{m}) = \sum_{n=N_1}^{N_2} \lambda_n(m_n) \quad . \tag{4.16}$$

It is seen that $\Lambda_{\underset{\sim}{m}}$ is a function of the vector $\underset{\sim}{m}$ through a sum of the functions of its coordinates.

This observation forms the basis for the application of the Viterbi algorithm to the solution of this demodulation problem. In fact, the hypothesis that the process $(\xi_n)$ forms a Markov chain means in particular that the set of values that each $\xi_n$ is allowed to take depends on the future values $\xi_{n+1},\xi_{n+2},\ldots$ only through $\xi_{n+1}$. Since we have denoted these values by $m_n$, this is equivalent to saying that the allowable values for $m_n$ will depend on the value of the other components $m_{n+1},m_{n+2},\ldots$ of $\underset{\sim}{m}$ only through $m_{n+1}$.

This assumption is crucial in order to allow the Viterbi algorithm to be applied to the solution of this maximization problem.

To see why this is true, consider for notational simplicity the case $N_1=1$, $N_2=N$. Then, the demodulation problem is equivalent to finding

$$\mu = \max_{m_1,\ldots,m_N} \sum_{n=1}^{N} \lambda_n(m_n) \tag{4.17}$$

i.e., maximizing a function of $N$ arguments made up of the sum of $N$ functions, each of them dependent on only one of the arguments.

Denote by

$$m_i \longrightarrow (m_{i+1},m_{i+2},\ldots) \tag{4.18}$$

the values that $m_i$ is allowed to take under the constraint that the following components of $\underset{\sim}{m}$ take values $m_{i+1},m_{i+2},\ldots$ . The maximization problem can be solved sequentially as follows:

$$\mu = \max_{m_N} \max_{m_{N-1}\to m_N} \cdots \max_{m_2\to(m_3,\ldots,m_N)} \max_{m_1\to(m_2,m_3,\ldots,m_N)} \sum_{n=1}^{N} \lambda_n(m_n) =$$

$$= \max_{m_N}\left\{ \max_{m_{N-1}\to m_N}\left\{ \cdots \max_{m_2\to(m_3,\ldots,m_N)}\left\{ \max_{m_1\to(m_2,\ldots,m_N)} \left\{\lambda_1(m_1)\right\} + \lambda_2(m_2)\right\} + \cdots \right.\right.$$

$$\left.\left. \cdots\right\} + \lambda_N(m_N)\right\} \tag{4.19}$$

This can be written recursively as follows:

$$\mu_2(m_2,\ldots,m_N) \overset{\Delta}{=} \max_{m_1\to(m_2,\ldots,m_N)} \lambda_1(m_1)$$

$$\mu_3(m_3,\ldots,m_N) \overset{\Delta}{=} \max_{m_2\to(m_3,\ldots,m_N)} \left\{\mu_2(m_2,\ldots,m_N) + \lambda_2(m_2)\right\} \tag{4.20}$$

$$\mu_4(m_4,\ldots,m_N) \overset{\Delta}{=} \max_{m_3\to(m_4,\ldots,m_N)} \left\{\mu_3(m_3,\ldots,m_N) + \lambda_3(m_3)\right\}$$

$$\vdots$$

$$\mu_N(m_N) \overset{\Delta}{=} \max_{m_{N-1}\to m_N} \lambda_{N-1}(m_{N-1})$$

$$\mu = \max_{m_N} \mu_N(m_N)$$

In words, fix first $m_2,\ldots,m_N$ and maximize $\Lambda_{\underset{\sim}{m}}$ with respect to the values of $m_1$ that can lead to those values of $m_2,\ldots,m_N$. Due to the structure of $\Lambda_{\underset{\sim}{m}}$ , this is equivalent to maximizing $\lambda_1(m_1)$ alone. This must be done for all possible (N-1)-tuples $m_2,\ldots,m_N$ , and results in a function of $m_2,\ldots,m_N$ that we call $\mu_2$ .

Fix now $m_3,\ldots,m_N$ and take the maximum of $\Lambda_{\underset{\sim}{m}}$ , with $m_1$ equal to the value previously obtained, with respect to the values of $m_2$ that can lead to those $m_3,\ldots,m_N$ . This is equivalent to maximizing $\mu_2(m_2,\ldots,m_N) + \lambda_2(m_2)$ , since this is the only part of the function that does depend on $m_2$ . This results in a function of $m_3,\ldots,m_N$ that we call $\mu_3$ , and so on.

Simplifications of this basic algorithm are possible, depending on the structure of the range of the vector $\underset{\sim}{m}$ . The simplest possible case arises when, for all $i$ , we have

$$\{m_i : m_i \to (m_{i+1},m_{i+2},\ldots)\} = \{m_i\} \tag{4.21}$$

i.e., the values that $m_i$ can take do not depend on the values of $m_{i+1},m_{i+2},\ldots$ . In our decoding problem, this situation occurs when the random variables $(\xi_n)$ are independent. In this case, we have

$$\mu = \max_{m_N} \lambda_N(m_N) + \max_{m_{N-1}} \lambda_{N-1}(m_{N-1}) + \cdots + \max_{m_1} \lambda_1(m_1) \tag{4.22}$$

which corresponds to bit-by-bit decoding.

The second simplest case arises when, for all $i$ , we have

$$\{m_i : m_i \to (m_{i+1},m_{i+2},\ldots)\} = \{m_i : m_i \to m_{i+1}\} \tag{4.23}$$

In words, the values that $m_i$ can take depend only on the next coordinate.

This situation occurs when the random variables $(\xi_n)$ are a Markov chain, where (4.21) becomes

$$\begin{aligned}
\mu_2(m_2) &\overset{\Delta}{=} \max_{m_1 \to m_2} \lambda_1(m_1) \\
\mu_3(m_3) &\overset{\Delta}{=} \max_{m_2 \to m_3} \{\mu_2(m_2) + \lambda_2(m_2)\} \\
\mu_4(m_4) &\overset{\Delta}{=} \max_{m_3 \to m_4} \{\mu_3(m_3) + \lambda_3(m_3)\} \\
&\vdots \\
\mu_N(m_N) &\overset{\Delta}{=} \max_{m_{N-1} \to m_N} \{\mu_{N-1}(m_{N-1}) + \lambda_{N-1}(m_{N-1})\} \\
\mu &= \max_{m_N} \mu_N(m_N)
\end{aligned} \tag{4.24}$$

Notice that sometimes further simplifications can occur. For example, if $\mu_i(m_i)$ does not depend on $m_i$, i.e., if all the values of $m_i$ give rise to the same value for $\mu_i(m_i)$, the following iterations can be simplified taking advantage of the fact that $\mu_i$ is now a constant.

Recursions (4.24) are known as the Viterbi Algorithm (see, e.g., Viterbi and Omura, 1977). The performance of such an optimum receiver can also be evaluated. Upper bounds to the probability of an error event, or to the bit error probability, can be computed (see Mesiya et al., 1977, or Viterbi and Omura, 1977), depending on the set of distances

$$d^2(\underset{\sim}{m},\underset{\sim}{m}') \overset{\Delta}{=} \int_0^{KT} [y_{\underset{\sim}{m}}(t) - y_{\underset{\sim}{m}'}(t)]^2 dt \ . \tag{4.25}$$

## 4.5 A MARKOV CHAIN MODEL FOR THE NOISY CHANNEL

Assume that the channel output is fed into a demodulator, whose output is a sequence of symbols $(r_k)_{k=-\infty}^{\infty}$ . To account for possible soft-decision demodulation, we can assume that $r_k$ takes values in the set $\{1,2,\ldots,M'\}$ , where $M' \geq M$ , and $M$ is the number of values taken by the random variables $a_n$ , the information source outputs.

Thus, we can think of the system including encoder, modulator, continuous channel and demodulator as a discrete channel, with inputs $\{1,\ldots,B\}$ and outputs $\{1,\ldots,M'\}$ (see Chapter 1 for notations). This channel is generally not memoryless, due to the presence of the coder and to the effects of the continuous channel.

To characterize this discrete channel, we shall build a Markov chain model of it. This model is derived from that obtained in §4.2, with the only addition of noise. Recall that the model of §4.2 describes the channel output by the sequence of states $(\xi_n)_{n=-\infty}^{\infty}$ .

If the values taken by $\xi_n$ were perfectly known, then the demodulation process would entail no error. The presence of noise at the channel output implies that errors can be made. Clearly, the probability of making an error in the demodulation process will depend on the actual value of $\xi_n$ , as well as on the noise.

In particular, assuming for simplicity that the demodulator is memoryless, its operation will be described through the function

$$r_n = \delta(\xi_n, \nu_n) \qquad -\infty < n < \infty \tag{4.26}$$

where $\nu_n$ represents the effect of the noise on the demodulation when the channel state is $\xi_n$ .

Define now the error sequence $(\varepsilon_n)$ as follows:

$$\delta(\xi_n,\nu_n)=\delta(\xi_n,0)\oplus\varepsilon_n \tag{4.27}$$

where $\oplus$ denotes addition mod M".

As far as the effect of the noise statistics on this model is concerned, I shall assume that the random variable $\varepsilon_n$ is independent of $\varepsilon_j$, $j\neq n$, and of $\xi_j$, $j\neq n$.

Define now a two-dimensional, discrete-time random process as the sequence $(\xi_k,\varepsilon_k)_{k=-\infty}^{\infty}$. This process, under our hypotheses, is a Markov chain with M' × M" states. In fact, define

$$\mathscr{T}(k,k-1,k-2,\ldots)\overset{\Delta}{=}\Pr\{(\xi_k,\varepsilon_k)|(\xi_{k-1},\varepsilon_{k-1}),(\xi_{k-2},\varepsilon_{k-2}),\ldots\} =$$

$$=\Pr\{\xi_k,\varepsilon_k \mid \xi_{k-1},\xi_{k-2},\ldots,\varepsilon_{k-1},\varepsilon_{k-2},\ldots\} \tag{4.28}$$

Under our assumptions, the value taken by $\varepsilon_k$ will not depend on the values taken by $\varepsilon_{k-1},\varepsilon_{k-2},\ldots$ ; thus

$$\mathscr{T}(k,k-1,k-2,\ldots) = \Pr\{\xi_k,\varepsilon_k|\xi_{k-1},\xi_{k-2},\ldots\} =$$

$$=\Pr\{\xi_k|\xi_{k-1},\xi_{k-2},\ldots\}\Pr\{\varepsilon_k|\xi_k,\xi_{k-1},\xi_{k-2},\ldots\} \tag{4.29}$$

Since $(\xi_k)_{k=-\infty}^{\infty}$ is a Markov chain, we get

$$\mathscr{T}(k,k-1,k-2,\ldots)=\Pr\{\xi_k|\xi_{k-1}\}\Pr\{\varepsilon_k|\xi_k,\xi_{k-1},\ldots\} \tag{4.30}$$

and finally, since $\varepsilon_k$ depends only on the actual state $\xi_k$ ,

$$\mathscr{F}(k,k-1,k-2,\ldots) = \Pr\{\xi_k|\xi_{k-1}\}\,\Pr\{\varepsilon_k|\xi_k\} \qquad (4.31)$$

In (4.31), indices $k-2,k-3,\ldots$ do not appear, so the pair $(\xi_k,\varepsilon_k)$ is independent of $(\xi_{k-2},\varepsilon_{k-2}),(\xi_{k-3},\varepsilon_{k-3}),\ldots$ . This is equivalent to assuming that our process forms a Markov chain. The transition probabilities of this Markov chain are obtained by the previous computations:

$$\begin{aligned}\Pr\{(\xi_k,\varepsilon_k) = (i,e)|(\xi_{k-1},\varepsilon_{k-1}) = (j,e')\} &= \\ &= \Pr\{\xi_k=i|\xi_{k-1}=j\}\,\Pr\{\varepsilon_k=e|\xi_k=i\}\end{aligned} \qquad (4.32)$$

Let $\underset{\sim}{P}$ denote the matrix whose entries are

$$(\underset{\sim}{P})_{ij} \triangleq \Pr\{\xi_k=j|\xi_{k-1}=i\} \qquad i,j=1,\ldots,M' \qquad (4.33)$$

and let

$$q_{ie} \triangleq \Pr\{\varepsilon_k=e|\xi_k=i\} \qquad \begin{array}{l} i=1,\ldots,M' \\ e=1,\ldots,M'' \end{array} \qquad (4.34)$$

(probability that the error is $e$ when the channel output state is $i$ ).

Then the Markov chain so constructed nas transition matrix

$$Q = \left[\begin{array}{cccc} \underset{\sim}{P}_0 & \underset{\sim}{P}_1 & \cdots & \underset{\sim}{P}_{M'-1} \\ \underset{\sim}{P}_0 & \underset{\sim}{P}_1 & \cdots & \underset{\sim}{P}_{M'-1} \\ & - & - & - \\ \underset{\sim}{P}_0 & \underset{\sim}{P}_1 & & \underset{\sim}{P}_{M'-1} \end{array}\right] \Bigg\updownarrow\ M''\text{ times} \qquad (4.35)$$

where

$$\underset{\sim}{P}_{\ell} = \underset{\sim}{P} \cdot \mathit{diag}[q_{\ell_0}, q_{\ell_1}, \ldots, q_{\ell_{M'}}] \qquad (4.36)$$

<u>Example</u>

Consider for example a binary, non-coded transmission scheme with hard decision modulator (M"=2) and a channel with M'=4 . Let the matrix $\underset{\sim}{P}$ describing the channel behavior be

$$\underset{\sim}{P} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Let

$$q_i = P\{\varepsilon_k = 1 | \xi_k = i\} \qquad i=1,2,3,4$$

and

$$q_i^* = 1 - q_i = P\{\varepsilon_k = 0 | \xi_k = 1\} .$$

Then

$$\underset{\sim}{Q} = \frac{1}{2}\begin{bmatrix} q_0 & q_1 & 0 & 0 & q_0^* & q_1^* & 0 & 0 \\ 0 & 0 & q_2 & q_3 & 0 & 0 & q_2^* & q_3^* \\ q_0 & q_1 & 0 & 0 & q_0^* & q_1^* & 0 & 0 \\ 0 & 0 & q_2 & q_3 & 0 & 0 & q_2^* & q_3^* \\ q_0 & q_1 & 0 & 0 & q_0^* & q_1^* & 0 & 0 \\ 0 & 0 & q_2 & q_3 & 0 & 0 & q_2^* & q_3^* \\ q_0 & q_1 & 0 & 0 & q_0^* & q_1^* & 0 & 0 \\ 0 & 0 & q_2 & q_3 & 0 & 0 & q_2^* & q_3^* \end{bmatrix}$$

From this characterization of the noisy channel, it is possible to get a number of useful parameters, such as the burst length distribution, describing the behavior of the channel. These parameters are particularly useful when a code has to be designed for use on this channel, since the memoryless assumption cannot be accepted.

For a discussion of these parameters, and how to obtain them, see for example Fritchman (1967), Tsai (1969) or for the best reference on this topic, the book by Blokh, Popov and Turin (1971). The model considered in this section has been used by Tatebayashi et al. (1975) to compute the probability of $\lambda$ errors in a sequence of $\mu$ symbols.

## R:FERENCES

M. Ajmone, E. Biglieri and M. Elia (1978). "Power spectra of digital signals after nonlinearities with memory", to be published.

E.L. Blokh, O.V. Popov, V. Ya. Turin (1971). Models of Error Generation in Chan.els for the Transmission of Digital Information, Izdatel'stvo Svyaz', Moscow (in Russian).

B.D. Fritchman (1967). "A binary channel characterization using partitioned Markov chains", IEEE Trans. on Inform. Theory, vol.IT-13, n.2, p. 24 ff, April.

M.F. Mesiya, P.J. McLane and L.L. Campbell (1977). "Maximum likelihood sequence estimation of binary sequences transmitted over bandlimited nonlinear channels", IEEE Trans. on Commun., to be published.

J.K. Omura (1971). "Optimal receiver design for convolutional codes and channels with memory via control theoretical concepts", Inform.Sciences, vol.3, p.243 ff.

M. Tatebayashi, M. Kasahara and T. Namekawa (1975). "Characteristics of decoding error in discrete-memory channel", Electronics and Communications in Japan, Vol. 58-A, n.4, p. 16 ff.

S. Tsai (1969). "Markov characterization of the HF channel", IEEE Trans. on Commun.Technol., vol. COM-17, n.1, p. 24 ff, February.

A.J. Viterbi and J.K. Omura (1977). Digital Communication and Coding, to be published.

## APPENDIX A: SPECTRAL ANALYSIS OF NON-STATIONARY RANDOM PROCESSES

The problem of computing the spectrum of an energetic quantity $\Pi$ associated with a random process $x(t)$ can be stated as follows: we want to find a function $\mathscr{G}(\omega)$ such that the two following conditions hold:

(i) $$\Pi = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathscr{G}(\omega)d\omega$$

(ii) Let $x(t)$ be passed through a linear, time-invariant system with transfer function $H(\omega)$ . Denoting by $x'(t)$ the output of this system, the corresponding energetic quantity associated with $x'(t)$ , say $\Pi'$ , is given by

$$\Pi' = \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(\omega)|^2 \mathscr{G}(\omega)d\omega$$

If conditions (i) and (ii) are fulfilled, then $\mathscr{G}(\omega)$ is called the spectrum of $\Pi$ .

As an example, for a wide-sense stationary stochastic process $x(t)$ , we can define its average power as

$$\Pi \overset{\Delta}{=} E\{|x(t)|^2\} \quad .$$

If we define

$$\mathscr{G}(\omega) \overset{\Delta}{=} \mathscr{F}[E\{x(t+\tau)x^*(t)\}]$$

( $\mathscr{F}$ denotes Fourier transform) it is known that (i) and (ii) hold, so that $\mathscr{G}(\omega)$ can be called the average power spectrum of $x(t)$ .

We want to consider non-stationary random processes; the appropriate subclass of random process for which we can define the spectrum of a useful energetic quantity is that of <u>harmonizable processes</u> (Loève, 1963). Roughly speaking, a process is harmonizable if we can define its Fourier transform; for a precise definition, see (Loève,1963) or (Cambanis and Liu,1970).

These processes are a first-step generalization of wide-sense stationary stochastic processes. It has been proved (Cambanis and Liu,1970) that, under some very mild conditions, a random process obtained as an output of a linear system is harmonizable. The system may be randomly time-variant; the input process need not be stationary or even harmonizable.

For a harmonizable random process, it makes sense to look for the spectrum of the following energetic quantity:

$$\Pi = ME\{|x(t)|^2\} \qquad (A.1)$$

where M denotes time average. The spectrum of $\Pi$ can be obtained as follows: define first the (generalized) function

$$\Gamma(\omega_1,\omega_2) \overset{\Delta}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\{x(t+\tau)x^*(t)\}\, e^{-j[(\omega_1-\omega_2)t+\omega_1\tau]}\, dt\, d\tau \qquad (A.2)$$

and try to express it in this form:

$$\Gamma(\omega_1,\omega_2) = 2\pi\ \mathscr{P}(\omega_1)\delta(\omega_1-\omega_2) + \Gamma^c(\omega_1,\omega_2) \qquad (A.3)$$

where $\Gamma^c(\omega_1,\omega_2)$ does not include any line masses on the bisector of the plane $(\omega_1,\omega_2)$ .

Then it can be shown (Blanc-Lapierre and Fortet,1968) that $\mathscr{P}(\omega)$ is the spectrum of $\Pi$ as defined in (A.1) (in other words, (i) and (ii) hold for $\mathscr{P}(\omega)$ ).

It is often useful to express $\Gamma(\omega_1,\omega_2)$ in the following form:

$$\Gamma(\omega_1,\omega_2) = E\{X(\omega_1)X^*(\omega_2)\} \qquad (A.4)$$

where $X(\omega)$ is the Fourier transform of $x(t)$ .

Example 1 (stationary processes)

Let $x(t)$ be a harmonizable, wide-sense stationary process; define its autocorrelation function

$$R(\tau) \overset{\Delta}{=} E\{x(t+\tau)x^*(t)\} .$$

Then

$$\Gamma(\omega_1,\omega_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(\tau)\, e^{-j[(\omega_1-\omega_2)t+\omega_1\tau]}\, dt\, d\tau \quad .$$

Integrating first in $t$ , we get

$$\Gamma(\omega_1,\omega_2) = \int_{-\infty}^{\infty} R(\tau)\, e^{-j\omega_1\tau}\, d\tau \cdot 2\pi\delta(\omega_1-\omega_2) \quad .$$

This means that

$$\mathcal{G}(\omega) = \int_{-\infty}^{\infty} R(\tau)\, e^{-j\omega\tau}\, d\tau$$

is the average power spectrum, and $\Gamma^c \equiv 0$ , i.e., the function $\Gamma(\omega_1,\omega_2)$ for wide-sense stationary random processes reduces to a line mass on the bisector of the plane $(\omega_1,\omega_2)$ .

<u>Example 2</u> (cyclostationary random processes)

Let $x(t)$ be a harmonizable, wide-sense cyclostationary process (see, for example, Gardner and Franks,1975). We have

$$E\,x(t+\tau)x^*(t) = E\,x(t+\tau+T)x^*(t+T) \tag{A.5}$$

Equation (A.5) can be interpreted by saying that this average, when considered as a function of $t$, is periodic with period $T$. We can represent it as a Fourier series, i.e.:

$$E\{x(t+T)x^*(t)\} = \sum_{n=-\infty}^{\infty} c_n(\tau)\, e^{jn\Omega t} \quad , \ \Omega = \frac{2\pi}{T} \tag{A.6a}$$

where

$$c_n(\tau) = \frac{1}{T}\int_0^T E\{(t+\tau)x^*(t)\}\, e^{-jn\Omega t}\, dt \tag{A.6b}$$

Using (A.6), we get

$$\Gamma(\omega_1,\omega_2) = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c_n(\tau)\, e^{jn\Omega t}\, e^{-j[(\omega_1-\omega_2)t+\omega_1\tau]}\, dt\, d\tau$$

$$= 2\pi \sum_{n=-\infty}^{\infty} C_n(\omega_1)\delta(\omega_1-\omega_2-n\Omega) \tag{A.7}$$

where $C_n(\omega)$, $-\infty < n < \infty$, are the Fourier transforms of $c_n(\tau)$. The function $\Gamma(\omega_1,\omega_2)$ is thus made by linear masses located on lines parallel to the bisector $(\omega_1,\omega_2)$. Thus, using (A.3)

$$\mathcal{S}(\omega) = C_o(\omega) = \frac{1}{T}\int_0^T \int_{-\infty}^{\infty} E\{x(t+\tau)x^*(t)\}\, e^{-j\omega\tau}\, d\tau\, dt \tag{A.8}$$

and this formula, derived differently, has been often used for spectral analysis of cyclostationary processes (see, for example, Robinson <u>et</u> <u>al</u>.,1973).

## References

A. Blanc-Lapierre and R. Fortet (1968), Theory of Random Functions, vol. 2, (Gordon & Breach, Sc.Publ., New York).

S. Cambanis and B. Liu (1970). "On harmonizable stochastic processes", Information and Control, vol.17, p. 183 ff, 1970.

W.A. Gardner and L.E. Franks (1975). "Characterization of cyclostationary random signal processes", IEEE Trans.Inform.Theory, vol. IT-21, p. 1 ff, January.

M. Loève (1963). Probability Theory, (van Nostrand Co., Princeton, NJ, publ.)

G. Robinson, O. Shimbo and R. Fang (1973). "PSK signal power spread produced by memoryless nonlinear TWTs", COMSAT Tech.Rev., vol. 3, n.2, fall.

## APPENDIX B: COMPUTATION OF THE MATRIX SERIES $\underset{\sim}{\Lambda}(\omega)$

The problem we want to solve in this Appendix is the following: compute the matrix series

$$\underset{\sim}{\Lambda}(\omega) = \sum_{n=0}^{\infty} \underset{\sim}{A}^{n} e^{-jn\theta} \tag{B.1}$$

where

$$\underset{\sim}{A} = \underset{\sim}{P} - \underset{\sim}{P}^{\infty} \quad , \tag{B.2}$$

$\underset{\sim}{P}$ being the $M \times M$ transition probability matrix of a homogeneous, regular Markov chain, and

$$\theta = \omega T \quad . \tag{B.3}$$

It is known that a necessary and sufficient condition for the equality

$$\sum_{n=0}^{\infty} \underset{\sim}{R}^{n} = (\underset{\sim}{I} - \underset{\sim}{R})^{-1} \tag{B.4}$$

to hold is that all the eigenvalues of $\underset{\sim}{R}$ have magnitude less than 1 . It is easy to prove (see Cariolaro and Tronca, 1974) that this is the case for the matrix $\underset{\sim}{A}\, e^{-j\theta}$ , so we can write

$$\underset{\sim}{\Lambda}(\omega) = (\underset{\sim}{I} - \underset{\sim}{A}\, e^{-j\theta})^{-1} \tag{B.5}$$

Thus, the matrix series $\underset{\sim}{\Sigma}(\omega)$ can be computed, for each value of $\theta$ , by inverting a matrix. This procedure is computationally inefficient because, if we want to compute the power spectrum for several values of $\omega$ , we need as many matrix inversions as values of $\omega$ .

For a more efficient technique, write $\underset{\sim}{\Lambda}(\omega)$ in this form:

$$\underset{\sim}{\Lambda}(\omega) = \sum_{i=0}^{L-1} \beta_i(\theta)\underset{\sim}{A}^i \qquad \text{(B.6)}$$

This is possible since, under our hypothesis, $\underset{\sim}{\Lambda}(\omega)$ turns out to be an analytic function of the matrix $\underset{\sim}{A}$. Hence, it can be written as a polynomial in $\underset{\sim}{A}$. The coefficients will generally depend on $\theta$, and the minimum value of $L$ in (B.6) equals the degree of the minimum polynomial of $\underset{\sim}{A}$.

To find the coefficients $\beta_i(\theta)$, write the minimum polynomial of $A$ as

$$d(\lambda) = \sum_{j=0}^{L} \alpha_j \lambda^j \quad , \quad \alpha_L = 1 \ . \qquad \text{(B.7)}$$

Then observe that, for the definition of the minimum polynomial,

$$\sum_{i=0}^{L} \alpha_i \underset{\sim}{A}^i = \underset{\sim}{0} \qquad \text{(B.8)}$$

($\underset{\sim}{0}$ is the null matrix). Equate now the right-hand sides of (B.5) and (B.6):

$$\underset{\sim}{I} = (\underset{\sim}{I} - \underset{\sim}{A}\, e^{-j\theta}) \sum_{i=0}^{L-1} \beta_i(\theta)\underset{\sim}{A}^i \qquad \text{(B.9)}$$

From (B.9), taking (B.8) into account, we get, after some algebra:

$$\beta_h(\theta) = \frac{\sum_{\ell=1}^{L-h} \alpha_{\ell+h}\, e^{j\ell\theta}}{d(e^{j\theta})} \quad , \quad 0 \le h \le L-1 \qquad \text{(B.10)}$$

Eq.(B.6) can also be rearranged according to the powers of $e^{j\theta}$ , giving

$$\underset{\sim}{\Lambda}(\omega) = \frac{\sum_{\ell=1}^{L}\left(\sum_{h=0}^{L-\ell} \alpha_{\ell+h} A^{h}\right) e^{j\ell\theta}}{d(e^{j\theta})} \tag{B.11}$$

Notice that the matrices appearing in (B.11) as coefficients of $e^{j\ell\theta}$ , $1 \leq \ell \leq L$ , can be evaluated once and for all at the beginning of the computation.

It can be observed that the hypothesis that $d(\lambda)$ is the minimum polynomial of $\underset{\sim}{A}$ has never been used. Actually, every $d(\lambda)$ such that (B.8) holds can be used instead of the minimum polynomial. Of course, the minimum polynomial will give the most compact form for $\underset{\sim}{\Lambda}(\omega)$ and, hence, for the spectrum.

As an example, the use of the characteristic polynomial of $\underset{\sim}{A}$ leads to a simple computational algorithm, due to Faddeev and first applied to this problem by Cariolaro and Tronca (1974). According to this technique, $\underset{\sim}{\Lambda}(\omega)$ can be evaluated as follows:

$$\underset{\sim}{\Lambda}(\omega) = \frac{\underset{\sim}{B}(e^{j\theta})}{\Delta(e^{j\theta})} \tag{B.12}$$

where $\underset{\sim}{B}(\cdot)$ is an $M \times M$ matrix polynomial:

$$\underset{\sim}{B}(\lambda) = \underset{\sim}{I}\, e^{jM\theta} + \underset{\sim}{B}_1\, e^{j(M-1)\theta} + \dots + \underset{\sim}{B}_{M-1}\, e^{j\theta} \tag{B.13}$$

and $\Delta(\cdot)$ is the characteristic polynomial of $\underset{\sim}{A}$:

$$\Delta(e^{j\theta}) = e^{jM\theta} + \delta_1\, e^{j(M-1)\theta} + \dots + \delta_M \tag{B.14}$$

The polynomials $B(\cdot)$ and $\Delta(\cdot)$ can be computed simultaneously using the following algorithm (Gantmacher,1960); starting with

$$\underset{\sim}{B}_0 \overset{\Delta}{=} \underset{\sim}{I}$$

let

$$\underset{\sim}{C}_k = \underset{\sim}{A}\, \underset{\sim}{B}_{k-1}$$

$$\delta_k = -\frac{1}{k} \operatorname{tr} \underset{\sim}{C}_k$$

$$\underset{\sim}{B}_k = \underset{\sim}{C}_k + \delta_k \underset{\sim}{I}$$

for $k=1,2,\ldots,M$ . At the final step, we must have $\underset{\sim}{B}_N = \underset{\sim}{0}$ , the null matrix.

With this algorithm, the coefficients of the polynomials whose ratio gives $\underset{\sim}{\Lambda}(\omega)$ can be computed only once, giving an expression for the spectrum which can be computed for several values of $\omega$ with a limited computational effort.

## REFERENCES

G.L. Cariolaro and G.P. Tronca (1974), "Spectra of block coded digital signals", IEEE Trans. on Commun., vol. COM-22, n.10, p. 1555 ff, October.

F.R. Gantmacher (1960), Matrix Theory, vol. I, Chelsea, New York (publ).

## APPENDIX C: A SERIES EXPANSION FOR HARMONIZABLE RANDOM PROCESSES

In this Appendix, a heuristic derivation is presented of a result, due to Masry et al.(1968) and to Campbell(1968), in series representations of wide-sense stationary stochastic processes. Similar results have been obtained for harmonizable processes by Cambanis and Liu(1970), and for the more general class of weakly continuous processes by Cambanis and Masry(1971).

Let $x(t)$ be a wide-sense stationary process; we want to write it as

$$x(t) = \sum_n \xi_n s_n(t) \qquad -\infty<t<\infty \quad . \tag{C.1}$$

Define first the Fourier transform of the process:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\, dt \tag{C.2}$$

so that $x(t)$ can be represented as

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\, d\omega \tag{C.3}$$

and (see Appendix A)

$$E\{X(\omega_1)X^*(\omega_2)\} = 2\pi\, \mathcal{S}(\omega_1)\delta(\omega_1-\omega_2) \tag{C.4}$$

where $\mathcal{S}(\cdot)$ is the average power spectrum of $x(t)$ .

Consider now the Hilbert space $\mathcal{H}(\mathcal{S})$ of the functions $f(\omega)$ with norm

$$\|f\|^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} |f(\omega)|^2\, \mathcal{S}(\omega)d\omega \tag{C.5}$$

and scalar product

$$(f,g) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(\omega)g^*(\omega)\,\mathscr{S}(\omega)d\omega \quad . \tag{C.6}$$

Let $\{\psi_n\}$ be a complete, orthonormal set in $\mathscr{H}(\mathscr{S})$ . Since

$$\|e^{j\omega t}\|^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} |e^{j\omega t}|^2\,\mathscr{S}(\omega)d\omega < \infty \tag{C.7}$$

we get

$$e^{j\omega t} \quad \epsilon \quad \mathscr{H}(\mathscr{S}) \tag{C.8}$$

so that we can expand it in the form

$$e^{j\omega t} \quad = \sum_n s_n(t)\psi_n(\omega) \tag{C.9}$$

where

$$s_n(t) = (e^{j\omega t},\psi_n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{j\omega t}\ \psi_n^*(\omega)\,\mathscr{S}(\omega)d\omega \ . \tag{C.10}$$

Multiplying both sides of (C.9) by $\frac{1}{2\pi}X(\omega)$ and integrating, we get, using (C.3):

$$x(t) = \sum_n \xi_n\ s_n(t) \tag{C.11}$$

where

$$\xi_n \overset{\Delta}{=} \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\psi_n(\omega)d\omega \ . \tag{C.12}$$

The random variables $\xi_n$ are uncorrelated; in fact

$$E\,\xi_m\xi_n^* = \frac{1}{(2\pi)^2}\iint_{-\infty}^{\infty} E\{X(\omega_1)X^*(\omega_2)\}\ \psi_m(\omega_1)\psi_n^*(\omega_2)d\omega_1\ d\omega_2$$

$$= \frac{1}{2\pi}\iint_{-\infty}^{\infty} (\omega_1)\delta(\omega_1-\omega_2)\ \psi_m(\omega_1)\psi_n^*(\omega_2)d\omega_1\ d\omega_2$$

$$= (\psi_m,\psi_n) = \delta_{mn} \quad . \tag{C.13}$$

<u>Example</u>

Suppose that

$$\mathcal{G}(\omega) = \begin{cases} G_o & |\omega| < \Omega \\ 0 & \text{elsewhere} \end{cases}$$

The system

$$\psi_n(\omega) = \sqrt{\frac{\pi}{\Omega G_o}}\ e^{jn(\pi/\Omega)\omega} \qquad |\omega| < \Omega$$

is orthonormal and complete in $\mathcal{H}(\mathcal{G})$ .

We have

$$s_n(t) = \frac{1}{2\sqrt{\pi\Omega G_o}}\int_{-\Omega}^{\Omega} e^{j\omega(t-n(\pi/\Omega))}\ G_o\ d\omega$$

$$= \sqrt{\frac{G_o\Omega}{\pi}}\ \frac{\sin(\Omega t-n\pi)}{\Omega t-n\pi}$$

and consequently

$$u(t) = \sum_{n=-\infty}^{\infty} \xi_n \frac{\mathit{sin}(\Omega t - n\pi)}{\Omega t - n\pi}$$

where

$$E\, \overline{\xi}_n \overline{\xi}_m = \frac{G_o \Omega}{\pi} \quad \delta_{nm} \, .$$

## REFERENCES

S. Cambanis and B. Liu (1970), "On harmonizable stochastic processes", Information and Control, vol. 17, p. 183 ff.

S. Cambanis and E. Masry (1971), "On the representation of weakly continuous stochastic processes," Information Sciences, vol. 3, p. 277 ff.

L.L. Campbell (1969), "Series expansions for random processes", in Proc.Int.Symp. on Prob. and Inform. Theory, Lecture Notes in Mathematics, No. 89, (Springer, New York, publ) p. 77 ff.

E. Masry, B. Liu and K. Steiglitz (1968), "Series expansion of wide-sense stationary random processes", IEEE Trans. on Inform.Theory, vol. IT-14, p. 792 ff.

## APPENDIX D: INEQUALITIES BASED ON MOMENTS

### D.1 Statement and Solution of the Problem

Let X be a continuous random variable whose range [a,b] and whose first N moments

$$\mu_i \overset{\Delta}{=} E\{X^i\} \qquad i=1,2,\ldots,N$$

are known. In this Appendix, we shall consider the two following problems:

(i) Find sharp upper and lower bounds to the probability

$$\Pr\{X \le \lambda\}$$

where $\lambda \in (a,b)$ is a known quantity.

(ii) Find sharp upper and lower bounds to the average

$$E\{\Omega(X)\}$$

where $\Omega(\cdot)$ is a known function.

By "sharp" I mean that the bounds cannot be further improved; i.e., that random variables do exist that meet the bounds while having range [a,b] and moments $\mu_1,\ldots,\mu_N$. No attempt will be made to get closed forms for the results. Instead, numerical techniques will be sought that are general enough to handle a large class of situations and computationally suitable from the viewpoint of speed and accuracy.

Consider the set $\mathscr{F}(\mu_1,\ldots,\mu_N)$ of the distribution functions $F(\cdot)$ having range in [a,b] and first N moments

$$\mu_i \overset{\Delta}{=} \int_a^b x^i \, dF(x) \qquad i=1,\ldots,N \quad . \tag{D.1}$$

In general, there are many distribution functions having such a set of moments; actually, they form a closed, convex set in the space of all distribution functions. Some points in this space, however, play a key role in the solution of our problems.

Consider the subset $\mathcal{F}_\nu$ of piecewise constant distribution functions having a finite number $\nu$ of points of increase, say $(x_1, x_2, \ldots, x_\nu)$, and saltuses $(w_1, w_2, \ldots, w_\nu)$. For these functions,

$$F(x) = \sum_{i : x_i \leq x} w_i \tag{D.2}$$

Thus, for any distribution function $F \in \mathcal{F}_\nu \cap \mathcal{F}(\mu_1, \ldots, \mu_N)$, we get

$$\mu_i = \sum_{j=1}^{\nu} w_j x_j^{\,i} \tag{D.3}$$

Define now the function

$$\epsilon(x_i) = \begin{cases} 2 & a < x_i < b \\ 1 & x_i = a \quad \text{or} \quad x_i = b \end{cases} \tag{D.4}$$

Then the <u>index</u> of the distribution function $F \in \mathcal{F}_\nu$ is defined as

$$I_F = \sum_{i=1}^{\nu} \epsilon\,(x_i) \quad . \tag{D.5}$$

According to the index value, a distribution function $F$ belonging to $\mathcal{F}_\nu \cap \mathcal{F}(\mu_1, \ldots, \mu_N)$ will be called

<u>canonical</u> , if $I_F \leq N+2$

<u>principal</u> , if $I_F = N+1$ .

Two results due to Krein (1951) are the key to solving-both problems (i) and (ii) (see also Karlin and Studden (1966), and the references therein):

Proposition 1

For any N , there are only two principal distribution functions.

If N is odd, say N=2n-1 , their points of increase satisfy

$$a < x_1 < x_2 < \dots < x_n < b$$

and

$$a = x_1 < x_2 < \dots < x_n < x_{n+1} = b$$

respectively.

If N is even, say N=2n , their points of increase satisfy

$$a = x_1 < x_2 < \dots < x_{n+1} < b$$

and

$$a < x_1 < x_2 < \dots < x_n < x_{n+1} = b$$

respectively.

Proposition 2

If one of the $x_i$ is given in advance, $x_i \in (a,b)$ , then there exists a unique canonical distribution function including this point.

Consider now the solution to problems (i) and (ii);

Theorem 1 (Krein 1951)

If $\lambda$ is a point of continuity of the distribution of $X$ , then

$$\sum_{x_i < \lambda} w_i \le \Pr\{X \le \lambda\} \le \sum_{x_i \le \lambda} w_i \qquad (D.6)$$

where $x_i, w_i$ are the points of increase and the saltuses of the canonical distribution in $\mathscr{F}(\mu_1,\ldots,\mu_N)$ including the point $\lambda$ .

Theorem 2 (Krein 1951)

If $\Omega(t)$ has a continuous (N+1)-th derivative, and $\Omega^{(N+1)}(t)$ is everywhere in $[a,b]$ either concave or convex, then

$$\int_a^b \Omega(x)dF'(x) \le E\{\Omega(X)\} \le \int_a^b \Omega(x)dF''(x) \qquad (D.7)$$

where $F'(\cdot)$ and $F''(\cdot)$ are the principal distribution functions in $\mathscr{F}(\mu_1,\ldots,\mu_N)$.

It is most important to observe that $F'(\cdot)$ and $F''(\cdot)$ which give the smallest and greates values of $E\{\Omega(X)\}$ do not depend at all on the choice of the function $\Omega(t)$ , provided that the convexity requirement requested in the statement of the theorem is met.

Actually, a solution to the problem of finding sharp upper and lower bounds to $E\{\Omega(X)\}$ when $\Omega^{(N+1)}(\cdot)$ is neither convex nor concave can also be found, but it involves such a large amount of computations that it is unsuitable for applications.

## D.2 Computational Algorithms

Once the solution to both problems (i) and (ii) has been found, the problem that arises is to devise simple algorithms to obtain the principal and canonical distribution functions in $\mathcal{F}(\mu_1,\ldots,\mu_N)$. So far, only problem (ii) has yielded a solution, which I shall present in this section. The rationale behind it should prove useful also to the solution of problem (i), although some modifications are needed.

Essentially, deriving principal and canonical distribution functions is equivalent to finding three sets of numbers $\{x_i\}_{i=1}^{\nu}$ , $\{w_i\}_{i=1}^{\nu}$ , $\{z_\ell\}_{\ell=1}^{p}$ , $a \le x_i \le b$, $w_i \ge 0$ , $z_\ell \ge 0$ , such that, given the set of moments $\mu_1,\ldots,\mu_N$ , the following holds:

$$\mu_j = \sum_{i=1}^{\nu} w_i x_i^{\,j} + \sum_{\ell=1}^{p} z_\ell y_\ell^{\,j} \tag{D.8}$$

Specifically:

- for the principal distribution functions, $N$ odd:

$$\nu = \frac{N+1}{2} \; , \quad p=0 \tag{D.9}$$

and

$$\nu = \frac{N-1}{2} \; , \quad p=2 \quad (y_1=a \; , \; y_2=b) \; ; \tag{D.10}$$

- for the principal distribution functions, $N$ even:

$$\nu = \frac{N}{2} \; , \quad p=1 \quad (y_1=a) \tag{D.11}$$

and

$$\nu = \frac{N}{2} \; , \quad p=1 \quad (y_1=b) \quad ; \tag{D.12}$$

- for the canonical distribution function including point $\lambda$ :

$$\nu \le \frac{N+1}{2} , \quad p\ge 1 \quad (y_1=\lambda) . \tag{D.13}$$

This problem, as we shall soon see, is equivalent to the problem of finding approximate integration formulas with a given degree of exactness. In fact, consider the approximation

$$\int_a^b f(x)dF(x) \cong \int_a^b f(x)dF_{\nu,p} \tag{D.14}$$

where $F_{\nu,\rho}$ is a distribution function with $\nu+p$ points of increase, $p$ of which are fixed. Define the remainder

$$R_{\nu,p}(f) \triangleq \int_a^b f(x)dF(x) - \int_a^b f(x)dF_{\nu,p}(x) . \tag{D.15}$$

We say that (D.14) has <u>degree of exactness</u> $s$ if the remainder is zero for $f(x) = x^{\ell}$ , $\ell=0,1,\ldots,s$ , and is nonzero for $f(x) = x^{s+1}$ . In other words, (D.14) has degree of exactness $s$ if $F_{\nu,p}(\cdot)$ is a discrete distribution function whose first $s$ moments take the same value as the corresponding moments of $F(\cdot)$ .

Thus, the problem of finding the principal distribution function is equivalent to the problem of finding approximations of the form (D.14) to the original distribution function $F(\cdot)$ .

The degree of exactness of the approximation must of course be equal to N, the number of known moments of the distribution $F(\cdot)$ . We have the following result (see Krylov (1962), p.161):

Theorem 3

Let $\{x_1,\ldots,x_\nu,y_1,\ldots,y_p\}$ be the set of points of increase of $F_{\nu,p}(\cdot)$, and $\{y_1,\ldots,y_p\}$ be the set of fixed points. The maximum degree of exactness of (D.14) is

$$2\nu + p-1$$

and this is achieved only if

$$E\{p_\nu(X)\rho(X)X^\ell\} = 0 \qquad \ell=0,1,\ldots,\nu-1 \tag{D.16}$$

where

$$p_\nu(x) = (x-x_1)\ldots(x-x_\nu) \tag{D.17}$$

and

$$\rho(x) = (x-y_1)\ldots(x-y_p) \quad . \tag{D.18}$$

As we can see from (D.9)-(D.12), *the problem of finding principal distribution functions is equivalent to the problem of finding approximations of the form (D.14) with the maximum degree of exactness.*

Consider first the situation p=0 (no preassigned point of increase for the principal distribution function). Then the maximum degree of exactness of (D.14) is $2\nu-1$, and this is achieved only if

$$E\{p_\nu(X)X^\ell\} = 0 \qquad \ell=0,1,\ldots,\nu-1 \quad . \tag{D.19}$$

The numerical algorithm for finding the points $x_1,\ldots,x_\nu$ in this case has been developed by Golub and Welsch (1969).

Condition (D.19) is equivalent to saying that $p_\nu(\cdot)$ is the $\nu$-th term in a sequence of orthogonal polynomials

$$(p_i(\cdot))_{i=0}^{\nu} \quad , \qquad deg\ p_i(\cdot) = i \tag{D.20}$$

such that

$$E\{p_n(X)p_m(X)\} = \alpha_n^2 \delta_{nm} \quad , \qquad \alpha_n^2 > 0 \tag{D.21}$$

Its zeros are then the points of increase of the distribution $F_{\nu,0}$ .

These polynomials satisfy a three-term recurrence relationship of the form

$$p_j(x) = (a_j x + b_j)p_{j-1}(x) - c_j p_{j-2}(x) \qquad j=0,1,\ldots,\nu \tag{D.22}$$

with initial values

$$p_{-1}(x) \equiv 0$$

$$p_0\ (x) = 1$$

and $a_j > 0$ , $c_j > 0$ .

To construct these polynomials, rewrite system (D.22) as follows:

$$x\underset{\sim}{p}(x) = \underset{\sim}{T}\ \underset{\sim}{p}(x) + \frac{1}{a_\nu}\ p_\nu(x)\underset{\sim}{e}_\nu \tag{D.23}$$

where $\underset{\sim}{p}(x)$ is the column vector of polynomials $p_0(x),\ldots,p_{\nu-1}(x)$ , $\underset{\sim}{T}$ is a tridiagonal matrix with

$$\left.\begin{aligned} (\underset{\sim}{T})_{ii} &= -\ b_i/a_i && i=1,\ldots,\nu \\ (\underset{\sim}{T})_{i,i+1} &= \ 1/a_i && i=1,\ldots,\nu-1 \\ (\underset{\sim}{T})_{i+1,i} &= \ c_i/a_i && i=1,\ldots,\nu-1 \end{aligned}\right\} \tag{D.24}$$

and

$$\underset{\sim}{e}_\nu = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \qquad \text{(D.25)}$$

Thus, we can see that $p_\nu(x)$ has a zero at $x=x_i$ if and only if

$$x_i \, \underset{\sim}{p}(x_i) = \underset{\sim}{T}\, \underset{\sim}{p}(x_i) \qquad \text{(D.26)}$$

i.e., $x_i$ is an eigenvalue of the tridiagonal matrix $\underset{\sim}{T}$ .

Hence, the problem of computing the set of points of increase of $F_\nu$ is reduced to the problem of evaluating the eigenvalues of a tridiagonal matrix.

It can also be shown that, if $\underset{\sim}{p}(x_i)$ is the eigenvector of $T$ corresponding to the eigenvalue $x_i$ , and in addition $\alpha_n=1$ in (D.21), then the choice

$$w_i = \frac{1}{\underset{\sim}{p}^T(x_i)\underset{\sim}{p}(x_i)} \qquad i=1,\ldots,\nu \qquad \text{(D.27)}$$

will give the saltuses of $F_{\nu,0}$ that achieve the degree of exactness $2\nu-1$ . In conclusion, the solution of an eigenvalue + eigenvector problem will completely specify the principal distribution function $F_{\nu,0}$ .

The problem now is to construct the polynomials $\{p_i(x)\}_{i=0}^{\nu}$ that satisfy the orthogonality relationship (D.20). This can be done starting from the knowledge of the moments

$$\mu_j = E\{X^j\} \qquad j=1,\ldots,2\nu$$

and using a procedure due to Mysovskih (1968).

Let $\sigma_0(x),\ldots,\sigma_n(x)$ be $n+1$ linearly independent functions, and let $\underset{\sim}{M} = (m_{ij})$ be the $(n+1)\times(n+1)$ matrix with elements

$$m_{ij} = E\{\sigma_i(X)\sigma_j(X)\} \qquad i,j=0,1,\ldots,n \tag{D.28}$$

Clearly, $\underset{\sim}{M}$ is positive definite. Let

$$\underset{\sim}{M} = \underset{\sim}{R}'\underset{\sim}{R} \quad , \quad R=(r_{ij}) \tag{D.29}$$

be the Cholesky decomposition of $\underset{\sim}{M}$, where $\underset{\sim}{R}$ is upper-triangular and $\underset{\sim}{R}'$ lower-triangular.

If

$$\underset{\sim}{S} \triangleq \underset{\sim}{R}^{-1} \quad , \quad \underset{\sim}{S}=(s_{ij}) \tag{D.30}$$

($\underset{\sim}{S}$ an upper triangular matrix), then it can easily be proved that

$$\phi_j(x) = \sum_{k=0}^{j} s_{kj}\ \sigma_k(x) \qquad j=0,1,\ldots,n \tag{D.31}$$

form an orthonormal system, i.e.,

$$E\{\phi_i(X)\phi_j(X)\} = \delta_{ij} \quad . \tag{D.32}$$

In particular, if

$$\sigma_i(x) = x^i \tag{D.33}$$

then $\phi_j(x)$ are polynomials, exactly those we are looking for. In this case, the matrix $\underset{\sim}{M}$ is the Hankel matrix with elements

$$m_{ij} = E\{X^{i+j}\} = \mu_{i+j} \qquad i,j=0,\ldots,n \tag{D.34}$$

To build the polynomial $p_\nu(x)$, we need thus the moments $\mu_0=1$, $\mu_1,\ldots,\mu_{2n}$.

Actually, only $2n-1$ moments are needed to construct the set orthogonal polynomials that we are looking for. In fact, the role of $\mu_{2n}$ is just that of normalizing the system of orthogonal polynomials, and its value affects neither the points of increase nor the saltuses of $F_{\nu,0}(\cdot)$ (Gautschi 1970, p.256).

Consider now the case $p=1$. Suppose first that

$$y_1 = a \tag{D.31}$$

so that

$$\rho(x) = x-a \qquad a\le x\le b \quad . \tag{D.36}$$

To construct the approximation (D.14) corresponding to this case and having degree of exactness $2\nu$, we must find a polynomial $p_\nu(x)$ such that

$$E\{p_\nu(X)(X-a)X^\ell\} = 0 \qquad \ell=0,1,\ldots,\nu-1 \quad . \tag{D.37}$$

This problem will be solved by reducing it to the same problem solved previously. In fact, condition (D.37) can be rewritten as

$$\int_a^b p_\nu(x)x^\ell\, d\bar{F}(x) = 0 \qquad \ell=0,1,\ldots,\nu-1 \tag{D.38}$$

where $\bar{F}(\cdot)$ is a new distribution function such that

$$d\bar{F}(x) = \frac{x-a}{\mu_1-a}\, dF(x) \quad . \tag{D.39}$$

The moments of $\bar{F}(x)$ are thus given by

$$\bar{\mu}_\ell = \int_a^b x^\ell \frac{x-a}{\mu_1-a}\, dF(x) =$$

$$= \frac{\mu_{\ell+1} - a\mu_\ell}{\mu_1-a} \qquad \ell=0,1,\ldots \qquad . \qquad (D.40)$$

Using this set of modified moments, we are thus able to compute the polynomial $p_\nu(x)$, and consequently its zeros. These zeros give the locations of the points of increase of the distribution $F_{\nu,1}$ requested.

To get the saltuses of these points that allow (D.14) to attain its maximum degree of exactness, we can use (Krylov 1962, p.164):

$$w_i = \frac{\mu_1-a}{\rho(x_i)}\, \bar{w}_i \qquad i=1,\ldots,\nu \qquad (D.41)$$

where $\bar{w}_i$ are the saltuses obtained by using (D.27). To compute the saltus $z_1$ at $y_1=a$, we need only observe that the sum of the saltuses must be equal to $\mu_0 \equiv 1$. Thus, from (D.8),

$$z_1 = 1 - \sum_{i=1}^{\nu} w_i \quad . \qquad (D.42)$$

Similar results hold for $y_1=b$ ; in this case $\rho(x)= b-x$ and the procedure is the same. In particular, the modified set of moments is now

$$\bar{\mu}_\ell = \frac{b\mu_\ell - \mu_{\ell+1}}{b-\mu_1} \qquad \ell=0,1,\ldots \qquad (D.43)$$

Consider finally the case $p=2$ . Now

$$y_1 = a \quad , \quad y_2 = b \qquad (D.44)$$

and

$$\rho(x) = (x-a)(b-x) \qquad (D.45)$$

To construct the principal distribution function, we must find a polynomial $p_\nu(x)$ such that

$$E\{p_\nu(X)(X-a)(b-X)X^\ell\} = 0 \qquad \ell=0,1,\ldots \tag{D.46}$$

Defining the new distribution function

$$\overline{F}(x) = \frac{(x-a)(b-x)}{-\mu_2+(a+b)\mu_1-ab}\, dF(x) \quad ; \tag{D.47}$$

then condition (D.47) is equivalent to

$$\int_a^b p_\nu(x)x^\ell\, d\overline{F}(x) = 0 \tag{D.48}$$

*The polynomial* $p_\nu(x)$, *hence its roots, can be obtained using the* technique previously outlined. To get the corresponding saltuses, we can use

$$w_i = \frac{-\mu_2+(a+b)\mu_1-ab}{\rho(x_i)}\, \bar{w}_i \tag{D.49}$$

To compute $z_1$ and $z_2$ , we can observe that in particular, since the degree of exactness of (D.14) must be at least 1 ,

$$\left.\begin{aligned} 1 &= \int_a^b dF_\nu(x) = z_1+ z_2+ \sum_{i=1}^{\nu} w_i \\ 1 &= \int_a^b x dF_\nu(x) = z_1 a+ z_2 b+ \sum_{i=1}^{\nu} x_i w_i \end{aligned}\right\} \tag{D.50}$$

We have two equations in two unknowns that can be solved to give $z_1$ and $z_2$ .

Example

Let $a=-1$, $b=1$, $N=3$ and

$$\mu_1 = 0 , \quad \mu_2 = \sigma^2 , \quad \mu_3 = 0 \quad .$$

The principal distribution functions with this set of moments have points of increase

$$-1 < x_1 < x_2 < 1$$

and

$$-1 \leq x_1 < x_2 \leq x_3 = 1 ,$$

respectively.

Compute first the principal distribution function with points of increase internal to the interval $(-1,1)$ . We get the orthogonal polynomial

$$p_2(x) = x^2 - \sigma^2$$

which has roots

$$x_1 = -\sigma, \quad x_2 = \sigma \quad .$$

The corresponding saltuses are given by

$$w_1 = w_2 = 1/2 \quad .$$

To compute the other principal distribution function, evaluate first the modified set of moments corresponding to

$$\overline{F}(x) = \frac{1-x^2}{1-\sigma^2} \; F(x) \quad .$$

In this case, we get only one moment:

$$\bar{\mu}_1 = \frac{-\mu_3 + \mu_1}{1-\sigma^2} = 0 \quad .$$

The corresponding orthogonal polynomial has degree 1 , and unique root 0 . The saltus is then $\bar{w}_1=1$ and, using (D.49):

$$w_1 = 1-\sigma^2 \quad .$$

The saltuses corresponding to points ±1 can be obtained solving (D.50):

$$\begin{cases} 1 = z_1 + z_2 + (1-\sigma^2) \\ 0 = -z_1 + z_2 \end{cases}$$

which gives

$$z_1 = z_2 = \sigma^2/2 \quad .$$

One final comment is appropriate about the restriction involved in Theorem 2 with respect to the (N+1)-th derivative of $\Omega(t)$ . Should this condition not hold, one can resort to bounds that are not sharp, but can be easily computed.

In fact, if N is an odd number, the principal distribution function, all of whose points of increase are internal to [a,b] , gives the approximation

$$E\{\Omega(X)\} \cong \sum_{i=1}^{\nu_0} w_i \, \Omega(x_i) \quad , \qquad \nu_o = \frac{N+1}{2}$$

valid for all $\Omega(x_i)$, where the remainder is bounded by

$$|R_{\nu,0}(\Omega)| \leq \frac{1}{(2\nu_0)!\, \alpha^2_{\nu_0}} \max_{\xi} |\Omega^{(2\nu_0)}(\xi)|$$

and $\alpha^2_{\nu_0}$ is defined in (D.21) (for details, see Benedetto and Biglieri, 1975, and references therein).

## REFERENCES

S. Benedetto and E. Biglieri (1975). "A computational method for solving noise problems", in: J.K. Skwirzynski, ed., New Directions in Signal Processing in Communication and Control, Noordhoff Int.Publ., Leyden (Holland).

W. Gautschi (1970). "On the construction of Gaussian quadrature rules from modified moments", Math.Comp., vol. 24, p. 245 ff., April.

G.H. Golub and J.H. Welsch (1969). "Calculation of Gauss quadrature rules", Math.Comp., vol. 23 , p. 221 ff, April.

S. Karlin and W.J. Studden (1966). Tchebycheff Systems: with Applications in Analysis and Statistics, J. Wiley & Sons, Interscience Publ., New York.

M.G. Krein (1951). "The ideas of P.L. Čebyšev and A.A. Markov in the theory of limiting values of integrals and their further developments", Am.Math. Soc.Transl.,Ser. 2, vol. 12.

V.I. Krylov (1962). Approximate Calculation of Integrals, Macmillan, New York (publ).

I.P. Mysovskih (1968). "On the construction of cubature formulas with the smallest number of nodes", Dokl.Akad.Nauk SSSR, v. 178, p. 1252 ff.